DOCUMENT RESUME

ED 235 193 TM 830 585

AUTHOR Wilson, Mark
TITLE Adventures in Uncertainty: An Empirical Investigation of the Use of a Taylor's Series Approximation for the Assessment of Sampling Errors in Educational Research.
INSTITUTION Australian Council for Educational Research, Hawthorn.
REPORT NO ACER-OP-17; ISBN-0-85563-469-3
PUB DATE Apr 83
NOTE 120p.; Document may be marginally legible due to small print.
AVAILABLE FROM The Australian Council for Educational Research, Frederick St., Hawthorn, Victoria, Australia 3122.
PUB TYPE Reports - Research/Technical (143)

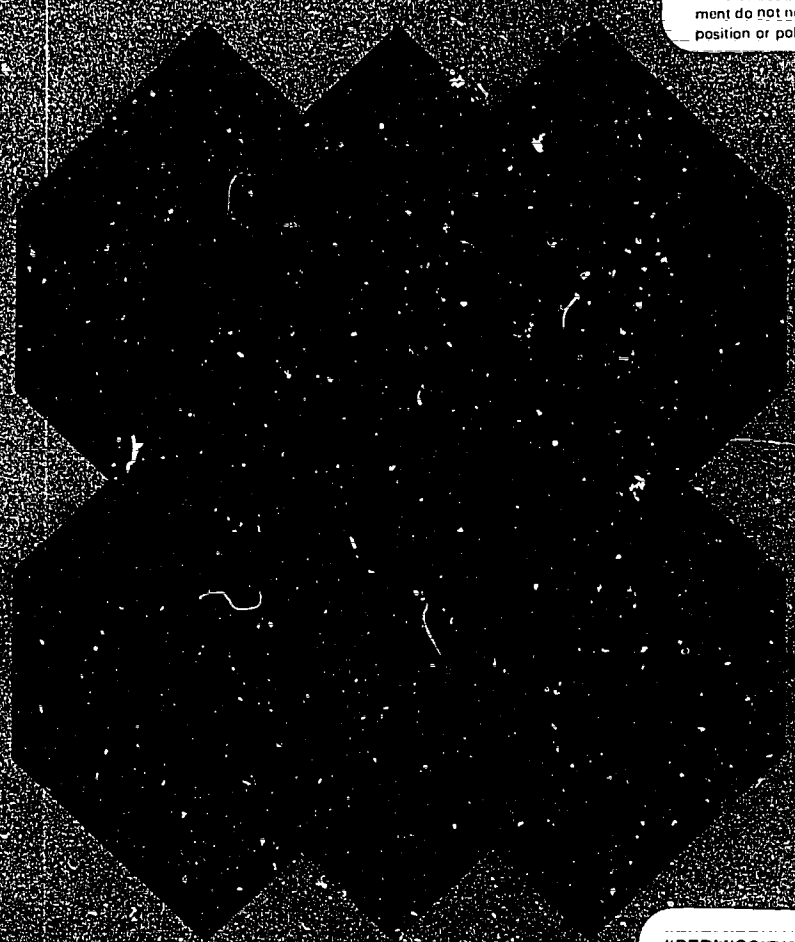
EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Data Collection; *Educational Research; Estimation (Mathematics); Guidelines; Mathematical Formulas; *Multivariate Analysis; Research Methodology; *Sampling; School Surveys; Statistics
IDENTIFIERS *Approximation (Statistics); FORTRAN Programing Language; *Sampling Error; Taylors Theorem

ABSTRACT

This study investigates the accuracy of the Woodruff-Causey technique for estimating sampling errors for complex statistics. The technique may be applied when data are collected by using multistage clustered samples. The technique was chosen for study because of its relevance to the correct use of multivariate analyses in educational survey research. To apply the technique the researcher must be able to write Fortran subroutines and must be able to ascertain a sampling error formula for a mean for whatever sampling situation is to be used (i.e. look up one of the standard texts). In return the technique will provide an estimate of the sampling error for any statistic which can be expressed in terms of a Fortran subroutine. Guides to numerical differentiation for the technique, and use of the technique and the writing of the Fortran subroutines are provided as appendixes to this paper. (PN)

Adventures in Uncertainty

Mark Wilson



ADVENTURES IN UNCERTAINTY

An empirical investigation of the use of a Taylor's series approximation for the assessment of sampling errors in educational research

Mark Wilson

Occasional Paper No. 17 - April 1983

Australian Council for Educational Research
Radford House, 9 Frederick Street,
Hawthorn, Victoria, Australia 3122

1983

Published by
The Australian Council for Educational Research,
Frederick St, Hawthorn, Victoria 3122.

Printed and bound by
Brown Prior Anderson,
5 Evans St, Burwood, Victoria 3125

National Library of Australia Cataloguing-in-Publication data

Wilson, Mark

Adventures in uncertainty.

Bibliography
ISBN 0 85563 469 3

1. Educational research. 2. Sampling (statistics)
3. Series, Taylor's. I. Australian Council for Educational
Research. II. Title. (Series: Occasional paper
(Australian Council for Educational Research); no. 17).

370'.7'8

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to Dr Kenneth N. Ross for his guidance and encouragement throughout this study. Through the many discussions we have had concerning the problem of variance estimation and by making available the data sets on which the evaluation was performed, he has greatly aided in the completion of this study.

To Dr J.P. Keeves, Director of the Australian Council for Educational Research, I must express many thanks for the use of the facilities of the Council for meeting many of the costs of the investigation. I would also like to thank my colleagues on the staff of the Council for their willingness to devote their time to solving many of the problems arising in the course of the investigation. In particular I would like to thank Mr J. Clancy and Mr S. Farish for their stalwart efforts to unravel the computational problems involved in the study.

I would also like to thank Ms J.S. Williams for her support in the practical aspects of thesis-writing.

Finally I am grateful to Ms Robyn Sperling, Mrs Carol Shackleton and Mrs Lynette Car for their skill and care in typing a manuscript which often resembled a jig-saw puzzle.

Mark Wilson

CHAPTER 1

INTRODUCTION

When an educational researcher conducts a survey it is almost always carried out in the administratively simple form of a clustered (and possibly weighted and stratified) sample of schools and classes. If the analysis of the data collected in this way is confined to means and differences between means, then the sampling variability, which is crucial to inference and to a complete understanding of the results, may be found using formulae available in the standard texts (for example, Cochran, 1963, and Kish, 1965). However, once the researcher attempts to use more sophisticated statistical procedures, the 'standard formulae' are found to apply only to simple random sampling. In the past, researchers have applied these erroneous 'standard formulae' and (hopefully) have handled the results with suspicion. Previous research (Peaker, 1975 and Ross, 1976) has shown that this suspicion is well-founded. The search for a solution to this problem has thrown up several approximate and intuitive techniques for estimating sampling errors given just one sample as evidence (Kish and Frankel, 1974). It is the purpose of this study to investigate the accuracy of one such approximation technique (Woodruff and Causey, 1976) under several of the types of sampling schemes that a typical educational research worker might be forced to employ.

Of course, the accuracy of the results is not the only criterion for evaluating such a technique. Ease of application is of great practical importance, as is flexibility in the face of the diverse statistical and sampling situations which arise in educational research. The particular technique to be studied was chosen because it was found to be the only technique available which struck a worthwhile balance between the demands it places on the skills of the research worker and the range of possible applications in which it would be suitable. To apply the technique the researcher must be able to write a few Fortran subroutines and must be able to ascertain a sampling error formula for a *mean* for whatever sampling situation is to be used (i.e. look up one of the standard texts). In return the technique will provide an estimate of the sampling error for *any* statistic which can be expressed in terms of a Fortran subroutine.

The two demands on the researcher are also investigated in this study. A guide to the use of the technique and the writing of the Fortran subroutines is provided as an Appendix in Microfiche to this Paper. Several approximation formulae for the sampling error of a mean, which might apply over a very wide range of sampling situations, are coupled with the technique and their performances evaluated. The establishment of an adequate approximation formula would considerably decrease the difficulty in applying the technique and open the way for its incorporation into 'user-oriented' packages.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The substance of most Sampling Theory textbooks (for example, Cochran (1963), and Kish (1965)), is the estimation of *descriptive statistics* and their standard errors for complex sample designs. Descriptive statistics are aggregates and means, and their ratios and products. However, many practitioners are also interested in estimating *analytical statistics* such as regression coefficients, discriminant functions and correlation coefficients, for the complex samples they use. Theory is lacking for the estimation of the standard error of analytical statistics for complex samples: researchers have been forced to resort to the formulae supplied by the textbooks for simple random sampling.

In order to alleviate this unfortunate situation attempts have been made to construct an appropriate theory with which to tackle the problem, but progress has been slow. Another solution was proposed by Tukey (1954):

> Statistical methods should be tailored to the real needs of the user ... 'What should be done' is almost always more important than 'what can be done exactly'. Hence new developments in experimental statistics are more likely to come in the form of approximate methods than in the form of exact ones.

Several techniques for approximating standard errors from single samples have been described: I shall refer to them collectively as 'single-sample techniques'.

There are the *replicated sampling* techniques of Jackknifing and Balanced Repeated Replication (also known as Pseudo-replication), the *random splitting* technique (also known as Independent Replication and Deming's Technique) and the *Taylor's series approximation* (variously known as the linearization method, the delta-technique, the propagation of error, and Taylorized deviations). A brief description of the first two and a more detailed analysis of the last follows.

2.2 Replicated Sampling Techniques

Replicated Sampling Techniques were first used by Mahalanobis (1944, 1946) in surveys of jute in India in 1936. Deming (1956, 1960) advocated designing samples which are easily broken-down into subsamples.

Two techniques which have gained prominence are *Balanced Repeated Replication* and the *Jackknife*.

Suppose that a statistic y is being used to estimate a parameter θ according to some sampling plan. The first technique, Balanced Repeated Replication, is used where the sample is divided into strata with two units selected from each stratum. The replication is a half-sample created by selecting one of the two sample units in each of the strata. The replication process is repeated g times. Then the estimates y_i' which are formed by estimating the parameter from the complementary half samples of the i^{th} replication may be used to approximate the variance of y thus:

$$\text{Var}(\bar{y}) \approx \frac{1}{g} \sum_{i=1}^{g} (y_i' - \bar{y})^2$$

McCarthy (1966) has shown that the most efficient strategy is to select orthogonal replications only.

For the second technique, the Jackknife, which was originally due to Quenouille (1956) and Tukey (1958), the sample is divided into g groups of size m. Then the values y_k, the estimates based on the m(g-1) observation remaining after deleting the k^{th} group of m observations, are used to ascertain the 'pseudovalues' y_k^* thus:

$$y_k^* = g\bar{y} - (g-1)\, y_k$$

These can then be used to form a jackknife estimate of θ

$$\bar{y}_b = \frac{1}{g} \sum_{k=1}^{g} y_k^*$$

and to estimate the variance of y

$$\text{Var}(y) \approx \text{Var}(\bar{y}_b) \approx \frac{1}{g(g-1)} \sum_{k=1}^{g} (y_k^* - \bar{y}_b)^2$$

Investigations by Miller (1968) have suggested that these estimates will be satisfactory when y can be expanded in a power series for each observation with

(i) the first-order term linear or regular in the observations;

(ii) second and higher-order terms negligible.

Similar, though less restrictive assumptions, will be made later for the Taylor's series approximation.

2.3 Random Splitting of Samples

The random subsample technique was developed by Deming (1960) following suggestions from J.W. Tukey. He estimated the variance of a statistic y by splitting the sample into 10 equal, independent and random subsamples, estimating the statistic for each subsample ($\bar{y}_i$) and for the entire sample ($\bar{y}$) and then approximating the variance of the statistic by the variance of the mean ($\bar{y}$) of the subsample statistics.

$$\text{Var}(\bar{y}) = \frac{\sum_{i=1}^{10} (\bar{y}_i - \bar{y})^2}{10(10 - 1)}$$

Ten was the number recommended by Tukey but the approximation holds, to a greater or lesser extent, no matter how many subsamples are taken.

Although superficially simple, this technique has several disadvantages for educational research. First, the estimation of complicated statistics may be neither stable, meaningful nor unbiased if only a small number of subsamples is taken (Finifter (1972), Mosteller and Tukey (1968)). Use of so many subsamples all modelled on the possibly clustered and stratified original sample would negate the computational simplicity of the original idea. Second, strata with small numbers of elements may need to be combined to allow the total sample to be divided into a large number of samples, resulting in a loss of detail. Third, if a large number of subsamples is used, outliers in the original sample will have little chance of appearing in some of the subsamples (Deming, 1956).

These difficulties have meant that researchers have concentrated on the other two techniques.

2.4 The Taylor's Series Approximation

The use of a Taylor's series approximation to obtain an estimate of the variance of a mean has been familiar to statisticians for some time. Its use for 'analytical statistics' was described by Deming (1960:390-396) and Kish (1965:585); and an early authoritative statement on its use was made by Kendall and Stuart (1963:231).

Let g be a function of the sample variates x_1, x_2 ... x_k, which are assumed to take the expected values θ_1, θ_2 ... θ_k. If g is differentiable at the point (θ_1, θ_2 ... θ_k), then the Taylor's series expansion of g about (θ_1, θ_2 ... θ_k) is

$$g(x_1, x_2 \ldots x_k) = g(\theta_1, \theta_2 \ldots \theta_k) + \sum_{i=1}^{k} \frac{\partial g}{\partial x_i} (x_i - \theta_i) \qquad (1)$$

$$+ \frac{1}{2!} \sum_{j=1}^{k} \sum_{i=1}^{k} \frac{\partial^2 g}{\partial x_1 \partial x_j} (x_i - \theta_i)(x_j - \theta_j)$$

$$+ \frac{1}{3!} \sum_{m=1}^{k} \sum_{j=1}^{k} \sum_{i=1}^{k} \frac{\partial^3 g}{\partial x_m \partial x_j \partial x_i} (x_m - \theta_m)$$

$$x \ (x_j - \theta_j)(x_i - \theta_i)$$

$$+ \ldots\ldots \text{ (Kendall \& Stuart, 1963 : 231-232)}$$

where the partial derivatives are calculated at the appropriate expected values. The first-order approximation to g is

$$Lg(x_1, x_2 \ldots x_k) = g(\theta_1, \theta_2 \ldots \theta_k) + \sum_{i=1}^{k} \frac{\partial g}{\partial x_i} (x_i - \theta_i) \qquad (2)$$

The first assumption made in the use of the Taylor's series approximation is that the sampling distribution of g is approximately equal to the sampling distribution of this linearized version of g. Thus

$$\begin{aligned} \text{Var}(g) &= \text{Var}(Lg) \\ &= \text{Var}\left(g(\theta_1, \theta_2 \ldots \theta_k) + \sum_{i=1}^{k} \frac{\partial g}{\partial x_i} (x_i - \theta_i)\right) \\ &= \text{Var}\left(\sum_{i=1}^{k} \frac{\partial g}{\partial x_i} x_i\right) \qquad (3) \end{aligned}$$

since $g(\theta_1, \theta_2 \ldots \theta_k)$ and $\sum_{i=1}^{k} \frac{\partial g}{\partial x_i} \theta_i$ are both constants (Frankel, 1971:28).

Actually using such an estimator depends of course upon obtaining values for the partial derivatives. The second assumption involved in the use of the Taylor's series approximation is that values of these partial derivatives obtained from the sample are reasonable approximations of their true values. Tepping (1968) made use of such a technique when he estimated the sampling variance of a regression coefficient over a multi-stage sampling design. Formulae for these partial derivatives are available for some of the more common statistics such as ratio means,

correlation coefficients, and regression coefficients (Frankel, 1971:30-31). However, beyond this the ground is as yet unexplored. Furthermore, although Tepping found a means of using equation (3) in the particular sampling situation he was investigating, he also noted that:

> ... the manner in which the variance of that linear approximation may be estimated will of course depend on the sample design. (Tukey, 1954:723)

Unfortunately the procedure for doing so is far from routine.

It was to this latter problem that Woodruff (1971) turned his attention. By restricting the variates to those which are sums of the observations (or sums of transformations of the observations), equation (3) may be re-expressed thus:

$$\text{Var}(g) = \text{Var}\left(\sum_{i=1}^{k} \frac{\partial g}{\partial x_i} \sum_{j=1}^{n} x_{ij}\right) \tag{4}$$

when it is assumed that the observed values have been enumerated from 1 to n for each variate x_i. As the two summations are finite, their order may be reversed to give

$$\text{Var}(g) = \text{Var}\left(\sum_{j=1}^{n} \sum_{i=1}^{k} \frac{\partial g}{\partial x_i} x_{ij}\right) \tag{5}$$

By defining a 'U - statistic' for each case by

$$U_j = \sum_{i=1}^{k} \frac{\partial g}{\partial x_i} x_{ij} \qquad j = 1,2 \ldots n \tag{6}$$

the equation becomes

$$\text{Var}(g) = \text{Var} \sum_{j=1}^{n} U_j \tag{7}$$

Now these U-statistics are simply univariate statistics which are linearly related to the original variates $x_1 \ldots x_k$. The formula for the evaluation of the variance in equation (7) is the one which would be appropriate for the estimation of the variance of a variable under the particular sampling design being used. This information is available in the standard texts for a wide range of sample designs (see, for example, Cochran (1963) and Hansen, Hurwitz and Madow (1953)). It should be noted that these standard texts will often quote a formula for the sampling error of the mean of a variable which will have to be adjusted to give the variance of the variable which is needed here. This procedure will be referred to as the Woodruff algorithm, or the Woodruff technique.

A Worked Example. The description of the algorithm used to estimate sampling errors will be made clearer and more concrete by the following example.

Consider the simple linear regression of the variable x on the variable y, with n observations,

$$y_i = a + bx_i + e_i \qquad i = 1 \ldots n \tag{8}$$

With the usual assumptions the best estimator of the regression slope b is

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{9}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

If we define,

$$s_i = (x_i - \bar{x})y_i \qquad i = 1 \ldots n \tag{10}$$

$$t_i = (x_i - \bar{x})^2 \qquad i = 1 \ldots n$$

as the variates to be used in the algorithm, then

$$\hat{b} = \frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} t_i} \tag{11}$$

and if

$$s = \sum_{i=1}^{n} s_i \tag{12}$$

$$t = \sum_{i=1}^{n} t_i$$

then

$$\hat{b} = \frac{s}{t} \tag{13}$$

Now the derivatives of the estimator with respect to each of the totals s and t may be found,

$$\frac{\partial \hat{b}}{\partial s} = \frac{1}{t}$$

$$\frac{\partial \hat{b}}{\partial t} = \frac{-s}{t^2} \qquad (14)$$

Recourse to equation (7) then gives the Taylor Approximation of the Variance of $\hat{b}$ as,

$$\text{Var}(\hat{b}) = \text{Var}\ \frac{\partial \hat{b}}{\partial s}\left(\sum_{i=1}^{n} s_i\right) + \frac{\partial \hat{b}}{\partial t}\sum_{i=1}^{n} t_i \qquad (15)$$

$$= \text{Var}\ \sum_{i=1}^{n}\left(\frac{s_i}{t} - \frac{st_i}{t^2}\right) \qquad (16)$$

$$= \text{Var}\ \sum_{i=1}^{n} U_i \qquad (17)$$

where $$U_i = \frac{s_i}{t} - \frac{st_i}{t^2} \qquad (18)$$

The variance involved in equation (17) is the variance appropriate for a total according to the particular sampling technique used.

The restriction to functions of statistics which are totals of the observations is not so great as it may appear at first glance. For instance, a statistic as complicated as a multiple correlation coefficient may be expressed as a function of the sums and sums of squares and sums of cross-products of the variables involved in the regression equation. In this case the original list of variates need only contain all of these in order that the Woodruff algorithm be implemented.

In a paper, written by Woodruff and Causey (1976), a computer program is described which implemented this algorithm and solved the problem of evaluating the partial derivatives by the use of a numerical technique which avoids the necessity of supplying a formula. It does however, involve the writing of at least one Fortran subroutine.

They checked the accuracy of this further approximation in three ways. First they compared the true partial derivatives with the numerical approximations, and found that the greatest relative difference was less than two parts in a million over a range of partial derivatives involved in the calculation of 48 different estimates in a six stratum sampling design.

Second, they compared the variance estimates for these 40 statistics given by the Taylor's series approximation using analytic derivatives; the relative differences were all less than one part in a million. Third, for those statistics for which no analytic derivatives were available, they compared the Taylor's series approximation using numerical derivatives with the Balanced Repeated Replication and Jackknife techniques over a very wide range of sampling designs; the results were found to be similar to those that Frankel (1971) achieved in comparisons using analytical derivatives.

2.5 Earlier Evaluations of the Taylor's Series Approximation

Several studies investigating the Taylor's series approximation for the estimation of standard errors were conducted without the Woodruff-Causey modifications. The three most important were those of Frankel (1971), Mellor (1973) and Bean (1975).

Frankel used data collected by the US Bureau of the Census in the 1967 Current Population Survey to simulate clustered stratified sampling on the basis of two primary sampling units per stratum. Comparison of the Replication, Jackknife and Taylor's series techniques was made for several sampling designs and for estimates of the mean, the difference of means, simple correlation coefficients, regression coefficients and multiple correlation coefficients. His conclusion was that although all three techniques gave satisfactory estimates of variance, (except possibly for the multiple correlation coefficient) the Taylor technique resulted in smaller mean square error whilst Balanced Repeated Replication gave a better approximation to Student's 't' statistic. Mellor's design was similar to this but used Monte Carlo simulation rather than existing population data and extended his comparison to partial correlation coefficients. His conclusions were essentially the same as those of Frankel, although he did note the comparative strength of Taylor's series approximation for error analysis of order statistics and highly skewed distributions. Bean, working at the National Center for Health Statistics, dismissed this use of synthetic populations as being 'of questionable representativeness'. She concluded from her study that both the Balanced Repeated Replication and the Taylor's series approximation gave adequate precision on the two criteria employed by Frankel, (Bean, 1975:10-14).

Accompanying the Woodruff-Causey paper was an empirical study using the same data as Frankel. The results of this study reinforced the conclusions of Frankel, although it was noted that the results using the Taylor's series approximation became substantially better with increased sample size. However Woodruff and Causey noted two other advantages of this technique.

> 1 The Taylor method is probably more economical for computer time, particularly in situations involving large numbers of strata (and/or simple draws). With the Taylor approximation, the basic data need be passed through the computer only twice, once to evaluate the partial derivatives and then again to form the substitute variables. The variances can then be computed with a single pass of these substitute variables. With the other two methods, the basic data must be tabulated a large number of times to obtain the results for a large number of partial samples. The 43,200 variances using the Taylor-N method for the 6, 12 and 30 strata designs required 38 minutes of UNIVAC 1108 central computer time (6 cents per variance at Census Bureau rates for this machine). The 21,600 variances for the 90, 270 and 810 strata designs required 85 minutes of UNIVAC 1108 central computer time (2.3 cents per variance). This includes the cost of the derivative evaluation as well as the actual variance computation.
>
> 2 The Taylor approximation is more versatile than the balanced replication method, and can easily be applied to any design for which there is a reasonable approximation to the variance of a single variable. The balanced replication method is most easily applied in sample designs involving a small number of strata and two draws per stratum. It can become difficult in other situations to find a balanced set of reasonable size. (Woodruff and Causey, 1976:321)

A recent survey by Shah (1978) recommended the Taylor's series approximation over the other three. He summarized the situation with Table 2.5.1 which is from his article. He also noted that whereas for the Taylor's series approximation the total cost of computing variances is about twice that of computing the mean only, the other techniques require between 50 and 100 times the cost of computing the mean. Furthermore he points out that if interpretation of the data requires the computation of variance components, the Taylor's series approximation is the only technique appropriate.

2.6 Some Theoretical and Practical Advances

Krewski and Rao (1978) have investigated the theoretical basis for the Taylor, Jackknife and Balanced Repeated Replication methods of sampling

Table 2.5.1 Comparison of Single Sample Techniques (from Shah 1978:32)

Criteria / Technique	Assumptions	Restrictions on sample design	Computational problems	Flexibility
Independent replications	Minimal	Severe	Simple	-
Pseudo-replication	Independence of complementary half replicates.	2 PSUs per stratum	Significant	-
Taylorized deviations	General central limit theorem	None	Not difficult	Can be used for variance components
Jackknife	Intuition	None	Greater than Taylorized deviation	Maybe *useful* for some designs

error estimation. They have established that as the number of strata approaches infinity, all three estimators are asymptotically normal and consistent. Although not very useful from a practical point of view, this result is nonetheless quite comforting. In a later paper they have also investigated the small sample properties of the three types of estimator; the results reported there are of interest but have a very restricted range of applicability due to the very strong model-assumptions necessary in such an investigation (Krewski and Rao, 1979).

Bobko and Reick (1980) have made an interesting application of the Taylor's series approximation to functions of correlation coefficients. As in equation (4) above they make the approximation of the function g of the correlation coefficients r_1, r_2 ,,, r_k thus,

$$\text{Var}\,(g(r_1, r_2 \ldots r_k)) = \sum_{i=1}^{k} [g_i'(\rho)]^2 \text{ var } (r_i) + \sum_{\substack{j=1 \\ j\neq 1}}^{k} \sum_{i=1}^{k} g_i'(\rho) g_j'(\rho) \text{ cov } (r_i, r_j)$$

where ρ is the expected value of the correlation coefficients, i.e.

$$\rho = (E(r_1), E(r_2) \ldots E(r_k))$$

Then using a normality assumption and some further restrictions they give formulae for $var(r_i)$ and $cov(r_i, r_j)$. Formulae for the derivatives are given for some simple statistics such as the correction for attenuation and indirect effects in path analysis. The resultant standard errors are then evaluated using data derived from synthetic populations. The emphasis on normal distributions points up the restriction in usefulness of this particular approach. In situations when the assumption of normal distributions was not tenable (which is often the reason for trying a Taylor Approximation) the expressions for the variance and covariance would not be applicable. The strength of this approach may lie not in the value of the actual standard errors obtained in any particular situation, but rather in the value of obtaining functional forms for the standard errors in terms of the correlation coefficients. The existence of such forms, even though based on quite restrictive assumptions, allows the investigation of sampling errors on a different level to that which has previously been possible.

Since the publication of the Woodruff-Causey paper, several general programs using Taylor series approximations have become available. There is, of course, the original Woodruff-Causey program. Next was Shah's STDERR (Standard Errors Program for Sample Survey Data) which computes certain ratio estimates or totals and their standard errors from the data collected in a complex multistage sample survey and is available within the SAS package (Shah, 1974). Hidiroglou, Fuller and Hickman (1975) published SUPER CARP (Cluster Analysis and Regressions Programme) which estimates totals, ratios, differences of ratios and regression coefficients and their associated variances for several multistage complex designs and for a one-fold nested error structure. M.M. Holt (1977) has produced SURREGR (Standard Errors of Regression Coefficients) for the testing of hypotheses concerning regression models using a stratified multistage sampling design and ordinary least squares or weighted least squares. The World Fertility Survey has produced a program called CLUSTERS which uses the 'collapsed strata' technique mentioned earlier to produce error estimates for ratio estimators (Verma and Pearce, 1978). The Office of Research and Statistics within the U.S. Social Security Administration is developing a software package designed to accommodate many different sampling designs but it is as yet able to offer the Taylor Approximation only in the Keyfitz form (see equation (5)) (Finch, 1978). A survey of

the many computer programs available, summarising a few important features for each, has also appeared (Kaplan, Francis and Sedransk, 1979). One method of evaluating these packages has been pursued by several researchers (Woodruff and Causey, 1976 and Maurer, Jones and Bryant, 1978). This involves the comparison of the programs with respect to their computational efficiency, evaluated in terms of central processing time, for a representative sample of designs. This comparison may loom large in the eyes of computer programmers, but for a research worker, the issues of ease of application and adaptability to different situations will prove much more important.

Although much valuable work has been done at many research centres, they have invariably been concerned with the solution of the sampling error problem in terms of the particular style of sample design dominant at each centre and in terms of the particular range of statistics that are studied there. The incorporation of sampling error routines into such packages as SAS and OSIRIS has begun and will eventually make the calculation of sampling errors a routine procedure within the limitations of the application of those packages. It would seem however that beyond this the researcher will be forced either to write entire programs for whichever single-sample technique is chosen, or to write the type of semi-standard subroutines which are necessary to the application of the Woodruff-Causey program.

2.7 Use of Variance Estimation in Educational Research

Attention to the problem of variance estimation by educational and psychological researchers was urged by Marks (1947) in connection with a revision of the Stanford-Binet Scale.

> Ignoring the effects of cluster sampling on measures of sampling error has undoubtedly resulted in attaching importance to results which are statistically insignificant. (Marks, 1947:413)

He found that the standard errors as calculated by the simple random sampling formulae were underestimating the true standard errors by a factor of three. The first investigation of sampling errors for a large-scale educational survey was made by Peaker (1953). Standard errors were found to be underestimated by half in this case.

The whole topic was consolidated with the work of Kish who introduced the statistic 'Deff' (design effect) which is

> the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements. (Kish, 1965:258)

A useful modification of this is the 'design factor', abbreviated as 'deft', and equal to the square root of the design effect. (Verma et al, 1980)

Kish used balanced repeated replication to estimate Deff values from a sample of 2,200 tenth grade boys in American public schools (Bachman et al, 1967). Deff for sample means was found to be less than three and for correlation coefficients and ratios it was found to be about 2.3.

A modification of Deming's technique using the range of estimates provided by four independent samples was used by Peaker (1967) in an international study of mathematics achievement. He found Deff values of correlation coefficients ranging from 1.96 (in Japan where clusters of 10 students per school were selected) to 8.4 (in Scotland where 75 students per school were selected).

Keeves (1966) decomposed total variance due to classroom and variance due to students in what appears to be the first application of these techniques to Australian educational data. He also calculated Deff values of from 1.00 to 21.3 using a similar method to that of Peaker (1967).

Jackknife procedures were used by Peaker (1975) in an international study of achievement (Comber and Keeves, 1973) in which he found average Deff values of 6 for means, 2.5 for correlations and 2 for regression coefficients; the primary sampling unit used was the school.

Ross (1976) used an empirical approach to estimate Deff for several typical sample designs and statistics in common use. He found that the lowest values of Deff occurred for designs that used schools as the primary sampling unit and also for the more complex multivariate statistics. A comparison of these results with Balanced Repeated Replication and Jackknife estimates revealed that both techniques were performing reasonably well on the average. However he points out that individually estimates vary quite considerably from the empirically-derived results.

CHAPTER 3

DESIGN OF THE STUDY

3.1 Introduction

The chapter which follows describes the procedures used to examine the Taylor Approximation. A previous study is described in detail, as the data-base and subsequent empirical analyses provide a bench-mark against which the technique can be compared. The comparison is in two parts. Firstly, the Taylor Approximation is compared to the empirically-established 'true' estimates of variance. Secondly, it is compared with two other single-sample techniques which were investigated in the previous study.

3.2 A Previous Study

The present study capitalizes on data collected by Keeves (1971) and later analysed by Ross (1976).

The remainder of the section is devoted to a summary of this data-base, and the analyses to which it was subjected by Ross. Further details may be found in Ross (1976) and Keeves (1971).

The Data-base. The population under study consisted of 2354 Year 7 students in the Australian Capital Territory in August 1969. This was 95 per cent of all such students: data sets which so nearly encompass a genuine population are extremely rare in educational research.

The students came from three school 'systems'. System 1 is a collection of nine government schools with fifty-three Year 7 classes. System 2 is a collection of four Catholic schools with fifteen Year 7 classes. System 3 is a pair of independent schools with seven Year 7 classes.

Keeves gathered data on a large range of variables for this population. Five were selected by Ross for inclusion in a causal model; they were chosen to represent a wide range of types of variable, to provide a range of magnitudes of the intercorrelations between them, and to constitute a meaningful model of educational achievement. These variables are described in Table 3.2.1.

The Causal Model. The causal model used by Ross is an example of the 'Path Analysis' technique (Duncan, 1975). This technique and its application to a particular situation could be subjected to any number of criticisms,

Table 3.2.1 The Variables in the Causal Model

Variable name	Description
SEX	Coded on a two point scale with male = 1, female = 2.
FOCCUP	The occupation of the student's father coded on a six point occupational prestige scale (Broom et al, 1977).
LIKESCHL	A 17 item scale designed to measure student's attitude towards school.
EXPEDN	A seven point rating designed to measure the student's level of aspiration for further education.
MATHS	A test of 55 mathematics items each of which was scored: correct = 1, incorrect = 0.

However, the model is used in this study merely as an example of the type of correlational analysis widely used in educational research.

The model investigates the relative influences among the variables under the assumption of a certain ordering of causality:

1 Antecedent student characteristics influence
2 Attitudes toward school and these characteristics and attitudes influence
3 Aspirations towards further education and these characteristics, attitudes, and aspirations influence
4 Achievement in Mathematics.

These influences are measured by what are termed 'path coefficients' which may be shown to be equal to standardized regression coefficients (Kerlinger and Pedhazur, 1973:310-14). The first stage in this causal chain consists of variables for which it is assumed that causes outside the model completely determine variability. At each subsequent stage it is assumed that causality is unidirectional; that is, no variable can be both cause and effect of another. A residual variable is included at each stage to account for all other sources of variation (these are referred to by lower-case letters a, b, c, etc.). It is assumed that a residual variable is neither correlated with other residual variables nor with the variables in the model to which it is not attached.

The model is illustrated in Figure 3.2.1. In interpreting correlation coefficients and path coefficients associated with this figure it should be

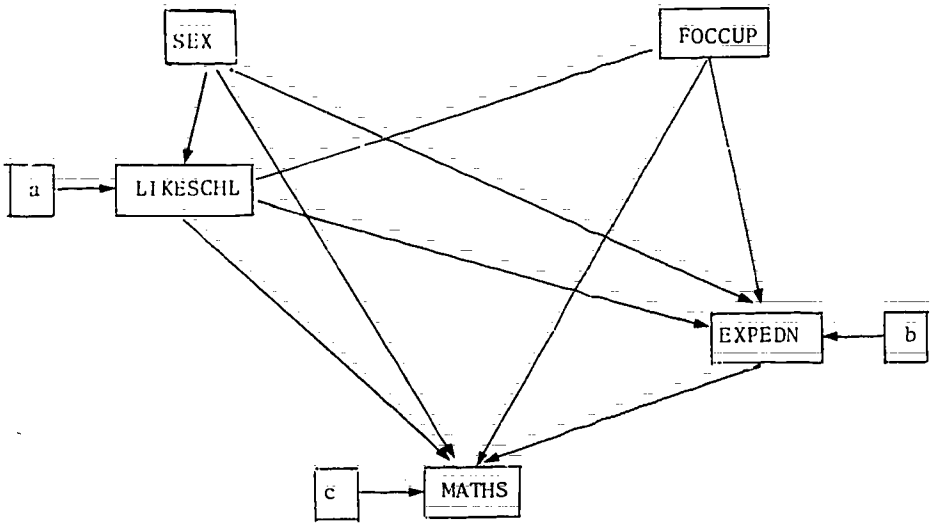


Figure 3.2.1 The Causal Model

noted that 'father's occupation' is not scored in the 'usual' direction, so that a high score on this variable assigns a low relative rating on the scale of occupational prestige.

The Sample Designs. In order to establish the effects of different sampling strategies, Ross chose five sample designs and drew, from the population described above, twenty-five samples of 150 students according to each of the sample designs. Samples of size 150 were deemed appropriate firstly as this is large enough to achieve stable estimates of the analytic statistics used in correlational analyses involving a 'medium' number of variables, and secondly as an example of the research designs which would be within the economic and administrative resources of the typical educational research worker. Twenty-five replications were considered sufficient to establish reliable empirical data for the sampling distributions of the various statistics associated with the causal model. The five sampling designs are described below.

Design 1: Simple random sample of 150 students (SRS design).

Each sample is a simple random sample of 150 students from the entire population.

Design 2: Stratified proportional simple random sample of 150 students (STR design).

Table 3.2.2 Contribution of Each Stratum to the STR Design

Stratum	Number of students in population	Proportion of population	Number of students in sample	Proportion of sample
1	1611	0.684	103	0.687
2	539	0.229	34	0.227
3	204	0.087	13	0.086
Total	2354	1.000	150	1.000

The strata chosen were the three school systems. Each stratum contributed to the sample in proportion to its size within the entire population; and within each stratum an independent simple random sample of students was chosen. The number of students from each stratum is shown in Table 3.2.2.

Design 3: Probability proportional to size selection of six primary sampling units (schools) followed by simple random selection of twenty-five students within each selected cluster (SCL design).

The fifteen schools were each allotted probability of selection according to their size, then six were chosen, without replacement, according to these probabilities. Within each school chosen, twenty-five students were selected as a simple random sample.

Design 4: Probability proportional to size selection of six primary sampling units (classes) followed by simple random selection of twenty-five students within each selected cluster (CLS design).

The sampling frame was first rearranged so that no class was smaller than twenty-five. Small classes were amalgamated to form 'pseudoclasses' and the same process was applied to these 'pseudoclasses' and to the larger classes as was applied to the schools in the SCL design.

Design 5: Stratified cluster sample of 150 elements with two primary sampling units (classes) being chosen from each stratum with probability proportional to size selection followed by simple random selection of 25 elements within each selected cluster (WTD design).

Table 3.2.3 Weights Used in the WTD Design

Stratum (h)	Number of students in the stratum (N_h)	Number of students in the sample (n_h)	Weight
1	1611	50	2.053
2	539	50	0.687
3	204	50	0.260
Total	N = 2354	n = 150	3.000

The sampling frame was first rearranged as for the CLS design. The same techniques were then applied to the set of classes and 'pseudoclasses' within each stratum as were applied to the SCL design, but only two selections were made. As this results in fifty students being selected from each stratum, the data for each student selected was weighted in proportion to the size of the stratum from which it was selected.

If N is the population size
n is the total sample size
N_h is the size of stratum h in the population
n_h is the size of stratum h in the population
then the weight for stratum h is

$$W_h = \frac{n}{N} \cdot \frac{N_h}{n_h}$$ (Kish, 1965:429)

Table 3.2.3 details the calculation of these weights for each stratum. (Note that for the other four sample designs each element of the population has the same chance of being selected, and hence, no weights were needed.)

The Sampling Error Formulae. The statistics chosen for study were: the Mean, the Correlation Coefficient, the Standardized Regression Coefficient and the Multiple Regression Coefficient. The sampling error formulae appropriate for each of these statistics under simple random sampling is given in Table 3.2.4. All except that for the Correlation Coefficient are standard results. For that statistic however, the more usual sampling error formula is

$$\hat{\sigma}_r = (1-r^2)/\sqrt{n}$$

Table 3.2.4 Sampling Error Estimation Formulae Used to Estimate the Denominator of the Equation which Defines the Design Effect

Sample statistic	Estimation formula	
Mean ($\bar{X}$)	$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$	(Guilford and Fruchter, 1973:127)
Correlation coefficient (r)	$\hat{\sigma}_r = \frac{1}{\sqrt{n}}$	(Guilford and Fruchter, 1973:145)[a]
Standardized regression coefficient (b)	$\hat{\sigma}_{12} = \frac{1-R^2_{1.234\ldots m}}{(1-R^2_{2.34\ldots m})(n-m)}$	(Guilford and Fruchter, 1973:368)
Multiple correlation coefficient (R)	$\hat{\sigma}_R = \frac{1}{\sqrt{n-m}}$	(Guilford and Fruchter, 1973:367)[b]

This formula was not used (a) because we wished to provide the reader with an example of how to use this technique in the relatively simple problem of testing whether the Correlation Coefficient is zero (for this test one assumes that r vanishes in order to calculate the sampling error and so the above formula reduces to the one given in Table 3.2.4) and (b) because there is some debate over the utility of this formula when r is small (See McNemar, 1969:155) which is the case for several of the correlations under investigation.

Results and Conclusions. Ross used the values of the square root of the Design Effect, 'deft' to measure the sampling errors. The equation defining this statistic is

$$\text{deft} = \frac{\hat{\sigma}_c}{\hat{\sigma}_{srs}}$$

where $\hat{\sigma}_c$ is the estimate of the standard deviation for the statistic and complex sampling design under consideration

and $\hat{\sigma}_{srs}$ is the estimate of the standard deviation for the same statistic which would be obtained if simple random sampling formulae were used.

The estimate $\hat{\sigma}_c$ was, of course, one of the goals of Ross's study. The formulae Ross used for $\hat{\sigma}_{srs}$ were derived from one source (Guilford and Fruchter, 1973), and are detailed in Table 3.2.4. The formulae apply to the case of a simple random sample of n elements on m variables where the variable X has a standard deviation of s. The multiple correlation coefficient $R_{1.23 \ldots m}$ refers to the regression equation which has variable 1 as the criterion and variables 2, 3, ... m as the predictors.

The values of the square root of the Design Effect (deft) for each statistic averaged over the twenty-five replications and for each of the five sample designs were calculated by Ross.

From this evidence Ross concluded that

> ... the use of complex sample designs to gather data may greatly influence the sampling stability of the statistics required to describe a recursive causal model. (Ross, 1976:45)

Ross also calculated the values of deft given by two of the single-sample techniques using one sample for each. Balanced Repeated Replication was used with the WTD design, and Jackknifing was used with the CLS design.

From these two cases Ross concluded that both techniques provided 'useful estimates of average $\sqrt{\text{Deff}}$'.

3.3 The Estimation of Sampling Errors from Single Samples Using a Taylor Approximation

The Woodruff Algorithm was applied to each of the five sample designs in order to estimate sampling errors. The process was repeated twenty-five times for each design to obtain a reliable guide to the behaviour of the estimate. The procedure followed is described in the remainder of this section, the results obtained are discussed in Chapter 4. Details of the application of the computer program may be found in Wilson (1981).

Variance Estimators for the Sample Designs. As mentioned previously, a Fortran subroutine providing an estimate of the variance of a total must be supplied to the program. The formula for each of the five sampling designs is given below. Let U be the statistic under consideration.

SRS Design

Let u_i be the observed value of the statistic for the i^{th} element
f be the total sampling fraction
n be the number of elements in the sample
$\bar{u}$ be the mean of the u_i's

Then the variance of the statistic U is estimated by

$$\text{Var}(U) = (1 - f)\frac{n}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2$$

(Woodruff and Causey, 1976)

STR Design

Let h=1,2 ... H be the strata

u_{hi}	be the observed value of the statistic for the i^{th} element in the h^{th} stratum
f_h	be the sampling fraction in the h^{th} stratum
n_h	be the number of elements sampled from the h^{th} stratum
$\bar{u}_h$	be the mean of the u_{hi}'s for the h^{th} stratum
W_h	be the proportion of the population in stratum h

Then the variance of the statistic U is estimated by

$$\text{Var}(U) = \sum_{h=1}^{H} W_h^2(1 - f_h)\frac{n_h}{n_h - 1}\left(\sum_{i=1}^{n_h}(u_{hi} - \bar{u}_h)^2\right)$$

(Woodruff and Causey, 1976)

SCL and CLS Designs

The most appropriate estimator for these two designs would be one which took into consideration the use of probability proportional to size selection and the use of selection without replacement at both stages of the two-stage design. Such an estimator is described by Sukhatme (1954: 410). However, this estimator involves the use of the probabilities of selection of the primary sampling units, and of the joint probabilities of selection of pairs of sampling units. This proved tractable though costly for the case of schools, but when the same computations were attempted for classes practical considerations involved in the use of busy computer installations meant that the job would never be finished. This problem is mentioned by Sukhatme who suggests that

> ... the use of the estimate appropriate for sampling with replacement, introducing the usual finite multiplier for calculating the error variance, is probably sufficiently satisfactory. (Sukhatme, 1954:415)

As this is the procedure most research workers would follow in any case it was decided to heed Sukhatme's advice.

As the two designs are exactly the same apart from the size of the clusters used, a unified account is given below.

Let m_i be the number of elements in the i^{th} cluster
m_o be the number of elements in the whole sample
n be the number of clusters sampled
f be the overall sampling fraction
$\bar{u}_i$ be the mean value of the statistic in the i^{th} cluster
$\bar{u}$ be the mean of the $\bar{u}_i$'s

Then the variance of the statistic U is estimated by

$$\text{Var}(U) = (1 - f)\frac{m_o^2}{n(n-1)} \sum_{i=1}^{n} (\bar{u}_i - \bar{u})^2$$

(Sukhatme, 1954:363)

WTD Design

For this design a version of the previous estimator could be applied within each stratum, the results weighted according to the relative sizes of the strata and then added across strata to obtain an estimate for the population variance. However, when this was attempted, the results proved extremely unstable due to the presence of only two clusters per stratum.
Two alternatives presented themselves, ignore the stratification and use the variance estimator for the CLS design, or ignore the clustering and use the variance estimator for the STR design. As the effects of clustering had already been investigated for two different designs, it was decided to pursue the latter strategy. Thus the variance estimator used was that described for the STR design, with the statistics u_{hi} and $\bar{u}_h$ replaced by the appropriate weighted statistics u'_{hi} and $\bar{u}'_h$.

CHAPTER 4

RESULTS: THE PERFORMANCE OF THE WOODRUFF TECHNIQUE FOR ESTIMATING SAMPLING ERRORS

In this chapter are discussed the performances of the Woodruff technique, as applied in the Woodruff-Causey program, as regards the estimation of sampling errors for the five designs and using the five variance estimators described in the previous chapter. The first section discusses the evaluation techniques used, the second examines the results for the SRS design, the third examines the results for the stratified designs and the fourth examines the results for the clustered designs. The fifth section compares these results with those obtained in a previous study, and the final section is a summary of these results.

4.1 The Evaluation Techniques

In discussing the effects of sample design, three types of evaluation procedures were used. The first measures the relationship between the estimates of sampling error obtained from the Woodruff-Causey program and the 'true' sampling errors which were derived empirically. The second type of evaluation relates to the internal consistency of the sampling error estimates which were obtained from the Woodruff-Causey program. The third type of evaluation investigates the extent to which the studentized ratios are distributed as a 't' - statistic around their mean, which bears upon their usefulness for hypothesis testing.

Design Effect. In order to establish a criterion for choosing between sample designs, Kish introduced the word 'Deff', derived from 'design effect', to name

> the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements (Kish, 1965:258).

Thus, if an estimator $\hat{u}$, of a population parameter u, is used under a complex sampling design C, then a measure of its efficiency is

$$\text{Deff } (\hat{u};C) = \frac{\text{Var } (\hat{u}_c)}{\text{Var } (\hat{u}_{srs})} \qquad (1)$$

Table 4.1.1 Population Values for the Statistics Used in the Study

Statistic		Population Value
Means:	SEX	1.4731
	FOCCUP	3.1175
	LIKESCHL	21.3732
	EXPEON	4.2840
	MATHS	29.5415
Correlation Coefficients:		
	SF	-0.01123
	SL	0.14908
	SE	-0.09723
	SM	-0.07560
	FL	-0.13988
	FE	-0.41609
	FM	-0.37256
	LE	-0.39518
	LM	0.21185
	EM	0.51094
Path Coefficients:		
	SL	0.14752
	SE	-0.15609
	SM	-0.04150
	FL	-0.13822
	FE	-0.36648
	FM	-0.19684
	LE	0.36719
	LM	0.02672
	EM	0.41444
Multiple Correlation Coefficients:		
	LIKESCHL	0.20329
	EXPEDN	0.55926
	MATHS	0.54211

where $\hat{u}_c$ indicates that the estimator is applied with the complex sample design C, whereas $\hat{u}_{srs}$ indicates that a simple random sample of the same size was used. Note, that Deff is dependent upon both the sample design and the estimator $\hat{u}$. Usually the relevant design and estimator are obvious and the arguments are left out.

As the discussion of the effects of sample design is usually couched in terms of sampling errors rather than sampling variance, a more appropriate criterion is the design factor or 'deft' which is defined by

$$\text{deft}(\hat{u},C) = \sqrt{\text{Deff}}(\hat{u},C) \qquad \text{(Verma et al, 1980)} \qquad (2)$$

Ross (1976) made all his comparisons using this measure. It has been pointed out that deft appears less sensitive to sampling errors than Deff (Kish, 1969:434).

In the interests of obtaining some stability in deft values, Kish and Frankel (1970:1092) recommend that particular values of deft be obtained for each instance of each type of statistic and that the average of these values should be reported as deft. Of course, such an averaging process must be confined to particular types of statistics due to differences in units of measurement, sample size, and differences in the variances of the variates involved in calculating the estimator.

The 'true' values of the various statistics were found using the SPSS collection of programs with 'list-wise' deletion. This means that the population parameters are slightly different to those quoted in Ross's study; this is not a problem as all the most important comparisons to be made were based on fresh samples. These values are given in Table 4.1.1 and will, for the purpose of this investigation, be considered true population parameters. The multiple correlation coefficients in this table are named by the criterion variable for the appropriate regression equation.

Calculation of design effects depends upon finding a good estimate of the standard error which would obtain under simple random sampling with the same number of sample cases as was used in the complex sample. The formulae used to calculate these simple random sampling standard errors were the same as those used by Ross (1976:29-30) which were detailed in Table 3.2.4.

In using these formulae, an estimate of the population standard deviation for each variable, and of the relevant Multiple Correlation Coefficients was found using the entire population. The formula standard error was then found using the appropriate number of sample cases.

This provides a 'best estimate' of the standard deviation that would be obtained from a simple random sample. In practice, a researcher would almost always have to use the sample obtained by the complex sampling process to estimate σ_{srs}. Such estimates would vary greatly depending on the particular sampling scheme in use. Use of the 'best estimate' provides a stable standard against which to compare both the empirical and the estimated standard errors of the complex designs.

These concepts were implemented according to the following formulae.

If f_i is the estimate of the function f resulting from the i^{th} sample, then the average, $\hat{f}$, is given by

$$\hat{f} = \frac{\sum_{i=1}^{25} f_i}{25} \tag{3}$$

and the *empirical* estimate of the standard deviation, $\hat{}_f$, is given by

$$\hat{s}_f = \sqrt{\frac{1}{25}\left(\sum_{i=1}^{25} f_i^2 - 25\hat{f}^2\right)} \tag{4}$$

Furthermore, if f is the 'true' value of the function (i.e. that derived from population data) and s_f is the simple random sample standard deviation derived from the formulae in Table 3.2.4, then the bias of $\hat{f}$ is given by

$$\text{bias } (\hat{f}) = \hat{f} - f \tag{5}$$

the Mean Square Error of $\hat{f}$ is given by

$$\text{M.S.E. } (\hat{f}) = [\text{bias } (\hat{f})]^2 + (\hat{s}_f)^2 \tag{6}$$

and an empirical estimate of deft is given by

$$\text{deft } (f_j, C) = \frac{\hat{s}_f}{s_f} \tag{7}$$

where C denotes the complex sampling design under consideration. In addition, the 'deft error' was also calculated; by this is meant, the percentage error incurred by assuming that deft equals one; that is,

$$\text{deft error} = \frac{1 - \text{deff}}{\text{deft}} \times 100 \tag{8}$$

Thus, a deft error of -26.43% indicates that if one used the simple random sampling version of the sampling error, one would be using an error estimate which was 26.43% *below* the correct figure.

Recourse to a table of the probability distribution of Student's 't' statistic on the appropriate number of degrees of freedom. (for example, Pearson and Wishart, 1947:118-119) will then provide an interpretation of this error in terms of true and apparent confidence intervals.

These values were then compared with the Woodruff-Causey estimated standard errors in the following way.

If s_{fi} is the i^{th} estimate of the standard error of function f, then the average standard error is simply

$$\bar{s}_f = \frac{\sum_{i=1}^{25} s_{fi}}{25} \tag{9}$$

the i^{th} estimate of deft is given by

$$\hat{deft}\ (f_i,C) = \frac{s_{fi}}{\bar{s}_f} \tag{10}$$

and the average deft is

$$\hat{deft}\ (f,C) = \frac{\sum_{i=1}^{25} \hat{deft}\ (f_i,C)}{25} \tag{11}$$

A percentage error involving this formula was also calculated using the formula

$$\text{deft error} = \frac{\hat{deft} - deft}{deft} \times 100 \tag{12}$$

where deft refers to the empirical value and $\hat{deft}$ refers to the average estimated deft for the function f.

Relative Mean Square Error. The internal consistency of the Woodruff-Causey estimates of standard error was investigated using the following statistics.

If s_{fi} is the i^{th} estimate of the standard error, and $\bar{s}_f$ is the average over the 25 samples, the standard deviation of the standard errors is given by

$$\text{st. dev. } (\bar{s}_f) = \sqrt{\frac{1}{25}\left(\sum_{i=1}^{25} (s_{fi})^2 - 25(\bar{s}_f)^2\right)} \tag{13}$$

the bias is given by

$$\text{bias } (\bar{s}_f) = \bar{s}_f - \hat{s}_f \tag{14}$$

and the Mean Square Error is given by

$$M.S.E.\ (\bar{s}_f) = [\text{bias}\ (\bar{s}_f)]^2 + [\text{st. dev.}\ (\bar{s}_f)]^2 \tag{15}$$

As these statistics gain meaningfulness only by comparison with the variability of the original function, f, and in order to allow comparison across function which have different magnitudes, the Relative Mean Square Error was also calculated.

$$\text{RELMSE}\ (\bar{s}_f) = \frac{\text{MSE}(\bar{s}_f)}{(\hat{s}_f)^2} \tag{16}$$

This can be broken down into two terms; Relative Bias, and Relative Variance given by

$$\text{RELBIAS}\ (\bar{s}_f) = \frac{[\text{bias}\ (\bar{s}_f)]^2}{(\hat{s}_f)^2} \tag{17}$$

$$\text{RELVAR}\ (\bar{s}_f) = \frac{[\text{st. dev.}\ (\bar{s}_f)]^2}{(\hat{s}_f)^2} \tag{18}$$

Of course, RELMSE = RELVAR + RELBIAS.

These statistics were those used by Frankel (1971:61-77), except that he investigated the variance rather than the standard deviation. In concordance with the use of 'deft' rather than 'Deff' it was decided that measures of the internal consistency of the standard error were more appropriate in this investigation.

Student's t. The third type of evaluation also follows the lead given by Frankel (1971). There he examined the assumption:

> The distribution of the ratio of the first-order estimate minus its expected value, to its estimated standard error is reasonably approximated by Students' t within symmetric intervals. (Frankel, 1971:78).

This assumption is crucial to the interpretation of the sampling errors derived from the Woodruff-Causey program. If the assumption is tenable, then credible inferences using the t-distribution can be made from the samples; if the assumption is not tenable, then the standard errors could still be utilized in a Tchebytchev - type inequality, but such results would be extremely conservative.

Table 4.1.2 Proportion of Students' t Within Selected Intervals

Degrees of Freedom	±2.576	±1.960	±1.645
3	0.9196	0.8549	0.8124
4	0.9384	0.8782	0.8244
5	0.9503	0.8925	0.8502
∞ (Standard normal case)	0.9900	0.9500	0.9000

Note: These proportions were (where necessary) calculated by linear interpolation from a table of the probability integral of Students' t in Pearson and Wishart (1947:118-119).

The investigation consisted of finding the proportion of times the ratio

$$\frac{\hat{f}_i - f}{s_{fi}} \qquad (19)$$

fell within the intervals (-2.576, 2.576), (-1.960, 1.960) and (-1.645, 1.645). These proportions were then compared to those predicted by 'Student's t' on an appropriate number of degrees of freedom.

Table 4.1.2 gives the Student's t proportions that were used for comparison.

For the non-stratified designs using a simple random sample variance estimator, the appropriate number of degrees of freedom is the number of sample cases minus one. Stratified designs usually take the number of cases minus the number of strata, but the presence of unequal stratum sizes and of weighting make this only an approximation. Frankel, investigating a series of designs involving many strata, but only two cases per stratum, hypothesized that the number of degrees of freedom was equal to the number of strata; this point is discussed in Section 4.4 (Frankel, 1971:79).

For the Jackknife variance estimator the appropriate number of degrees of freedom is one less than the number of distinct pseudovalues (Mosteller and Tukey, 1977:36). For all the Jackknife examples used in this study, there were six different pseudovalues, so the number of degrees of freedom was five. For the Balanced Repeated Replication variance estimator, the number of degrees of freedom was four. For the variance estimator used in the SCL and CLS designs the number of degrees of freedom is the number of clusters minus one; in this case, five.

Table 4.2.1 Average Deft Estimates for each Statistic

Average 'N' was 145.80

Function		Empirical	Estimated	Percent error of estimator	Percent error of formula
Means:	SEX	1.0057	0.9691	-3.6	-0.6
	FOCCUP	0.9423	0.9593	1.8	6.1
	LIKESCHL	1.1018	0.9610	-12.8	-9.2
	EXPEDN	1.0507	0.9842	-6.3	-4.8
	MATHS	0.8741	0.9629	10.2	14.4
Correlation Coefficients:					
	SF	0.9787	0.9667	-1.2	2.2
	SL	0.7320	0.9325	27.4	36.6
	SE	0.8421	0.9610	14.1	18.7
	SM	1.1150	0.9486	-14.9	-10.3
	FL	1.0966	0.9003	-17.9	-8.8
	FE	0.7693	0.8174	6.3	30.0
	FM	0.8022	0.8170	1.8	24.7
	LE	1.0302	0.8266	-19.8	-2.9
	LM	0.9649	0.8917	-7.6	3.6
	EM	0.7169	0.6826	-4.8	39.5
Path Coefficients:					
	SL	0.7524	0.9295	23.5	32.9
	SE	0.8634	0.9368	8.5	15.8
	SM	1.1406	0.9068	-20.5	-12.3
	FL	1.1320	0.9008	-20.4	-11.7
	FE	1.1493	0.9367	-18.5	-13.0
	FM	0.9582	0.9530	-0.5	4.4
	LE	1.1172	0.9695	-13.2	-10.5
	LM	1.0471	0.9413	-10.1	-4.5
	EM	0.9078	0.8836	-2.7	10.2
Multiple Correlation Coefficients:					
	LIKESCHL	0.8662	0.8852	2.2	15.5
	EXPEDN	0.6620	0.6371	-3.8	31.1
	MATHS	0.6756	0.6246	-7.5	48.0

Note: Values recorded in columns 1, 2 and 4 could be improved by making corrections for cases where $\bar{r} \neq 0$ or $\bar{R} \neq 0$.

4.2 Results for the SRS Design

The application of the Woodruff technique to this design may seem superfluous, after all, estimators for sampling errors for this design are well established. The investigation is important however, firstly because it provides a bench-mark against which to compare the results for all the other sample designs, and secondly because the 'formula' sampling errors quoted in Table 3.2.4 are all dependent upon some sort of normal-distribution assumption. This may not be appropriate. In addition, it should be noted that the formulae in Table 3.2.4 are appropriate for sampling with replacement from infinite normal populations. The methods used in this investigation relate to sampling from finite populations without replacement.

The average of deft for each of the statistics is given in Table 4.2.1. The first column gives the empirical values of deft obtained from the 25 simulations. The degree of variation from 1 indicates just how tenable was the 'formula' standard error; the empirical values range from 0.66 to 1.14 which indicates that the non-normal nature of the distributions of the variables is having considerable influence on the sampling errors of the statistics. The second column gives the estimated value of deft given by 25 applications of the Woodruff technique. It is striking that, except for the Multiple Regression Coefficients, the values in this column show much less variation than those in the previous column.

This in itself is not altogether a problem; if one is concerned primarily with the quality of the approximation for *each* of the statistics, it is worrisome. However, if the aim is to arrive at a reasonable deft estimate for each *type* of statistic, it need not be a problem at all. Kish and Frankel (1970:1092) recommend exactly this latter course, and in the main, their advice is hereby adhered to although in some cases comment is made on individual statistics. The third column gives the error in the deft estimate relative to the empirical deft. The worst error is 27% for the correlation between SEX and LIKESCHL. The final column gives the error involved in using the 'formula' version of sampling error (that is, assuming that deft is 1) relative to the empirical situation. The worst errors in this column are, for individual statistics, considerably worse than for the previous column.

The information contained in 4.2.1 is summarised by type of statistic in Table 4.2.2. The Woodruff technique is providing a slight underestimate of deft that is no more than 10 per cent in error. The 'formula' estimate

Table 4.2.2 Average Deft Estimates (SRS samples)

Average 'N' was 145.80

Statistic	Empirical	Estimated	Percent error of estimator	Percent error of formula
Means	0.9949	0.9673	-2.8	0.5
Correlation Coeff.	0.9048	0.8744	-3.4	10.5
Path Coefficients	1.0076	0.9287	-7.8	-0.7
Mult. Corr. Coeff.	0.7346	0.7157	-2.6	36.1

of deft is relatively better for the Means and the Path Coefficients and relatively poorer for the Correlation Coefficients and the Multiple Correlation Coefficients. One way of assessing the importance of these errors is to examine the real meaning that 95 per cent confidence intervals would have if these erroneous deft values were used. Table 4.2.3 gives the probability of an incorrect statement if a two-sided 95 per cent confidence interval is used: the probability should be 0.050. The 'formula' standard error for Multiple Correlation Coefficients is found to be very conservative, but all the rest would most probably be acceptable to most educational researchers.

Table 4.2.3 Probability of an Incorrect Statement About the Statistics in the SRS Design

	Probability of incorrect statement when a two-sided 95% confidence interval is to be used.	
Statistic	'formula'	Woodruff estimate
Means	0.049	0.057
Correlation Coefficients	0.030	0.058
Path Coefficients	0.052	0.069
Multiple Correlation Coefficients	0.008	0.056

Table 4.2.4 Average Bias and Variance Contributions to the Relative Mean Square Errors for the Statistics in the SRS Design

Statistic	Relative Bias	Relative Variance	Relative Mean Square Error
Means	0.006	0.002	0.008
Correlation Coefficients	0.020	0.004	0.025
Path Coefficients	0.023	0.004	0.028
Multiple Correlation Coefficients	0.002	0.012	0.015

Table 4.2.4 gives the contributions of the Bias and the Variance to RELMSE for the statistics under study. The variance contribution is very stable for the first three statistics, not ranging above one part in a hundred. Thus the variance estimator is about 1 per cent as variable as the statistic itself. For some individual statistics the Bias component is smaller than the Variance component, but on average, for all three types of statistics, the Variance component is *much* smaller than the Bias component.

For the Multiple Correlation Coefficient the situation is reversed with the performance of the estimator revealing quite a bit of variability, but on the average settling down to a good estimate. This contrary behaviour is echoed in the other designs.

The proportion of times that the 't' ratio falls within certain intervals for each type of statistic is given in Table 4.2.5. The appropriate number of degrees of freedom in 145 which is approximated by the entries for infinite degrees of freedom in Table 4.1.1. The results are tolerably close to the theoretically correct proportions except for the Path Coefficients which seem slightly more spread out than a true t-distribution.

Table 4.2.5 Proportion of Times that 't' Ratio Falls Within Selected Intervals (SRS samples)

Statistic	± 2.576	±1.960	±1.645
Means:	0.992	0.936	0.872
Correlation Coefficients	0.984	0.944	0.896
Path Coefficients	0.978	0.933	0.853
Multiple Correlation Coefficients	0.973	0.947	0.893

Table 4.3.1 Average Deft Estimates (SCL samples)
Average 'N' was 145.00

Statistic	Empirical	Estimated	Percent error of estimator
Means	1.4973	1.8113	21.0
Correlation Coefficients	1.0098	0.9676	-4.2
Path Coefficients	0.9998	1.0022	0.2
Multiple Correlation Coefficients	0.6782	2.6529	291.2

4.3 Results for the Clustered Designs: SCL and CLS

As these two designs are identical except for the relative sizes and nature of the clusters, their results are best considered together. Deft estimates for the two designs are listed in Tables 4.3.1 and 4.3.2: one is immediately struck by the huge overestimate for the Multiple Correlation Coefficients. The other statistics seem to be reasonably well estimated. The probabilities given in Table 4.3.3 are, except for the Multiple Correlation Coefficients just a little worse than those for the SRS design. Note that the basis of the calculation of these probabilities is different from that used for the previous design as there is now only five degrees of freedom involved in the variance estimation formula.

One way of considering these results is to calculate the 'effective sample size' for the two designs (Kish, 1965:259). This is the size of a simple random sample over the same variable which would give standard errors of the same size as were found here. Ross (1976:8) has given an approximate formula for the effective sample size in the case of the mean. If the population size is large compared to the sample size n, then the effective sample size n* is given by

$$n^* = \frac{n}{Deff}$$

Using this formula, the effective sample size for the Means in the SCL design is approximately 65, and for the CLS design, it is approximately 30. This certainly provides grounds for explaining the lowered performance of the Woodruff estimator in the case of Means. Unfortunately no such formula

Table 4.3.2 Average Deft Estimates (CLS samples)
Average 'N' was 144.96

Statistic	Empirical	Estimated	Percent error of estimator
Means	2.2068	2.4424	10.7
Correlation Coefficients	1.2173	1.1047	-9.3
Path Coefficients	1.1664	1.0739	-8.0
Multiple Correlation Coefficients	1.1080	3.4081	207.6

is available for the other statistics, although one might speculate that n* for the more complicated statistics would be closely related to n* for the means. If this is true then perhaps an explanation could be put forward for the poor behaviour of the estimator in the case of the Multiple Correlation Coefficients on the grounds that, with an effective sample size of 65 or 30, Multiple Correlation Coefficients themselves have little meaning or stability, and hence, the calculation of sampling errors is not warranted.

The bias and variance contributions to RELMSE are given in Table 4.3.4. The situation as for defts is reflected here: the results for the statistics other than Multiple Correlation Coefficients are reasonable but not so good as for the SRS design, and they are generally similar for both

Table 4.3.3 Probability of an Incorrect Statement About the Statistics in the SCL and CLS Designs

	Probability of incorrect statement when a two-sided 95% confidence interval is to be used	
Statistic	SCL design	CLS design
Means	0.027	0.036
Correlation Coefficients	0.057	0.067
Path Coefficients	0.050	0.064
Multiple Correlation Coefficients	<<0.001	<<0.001

Table 4.3.4 Bias and Variance Contributions to the Relative Mean Square Error for the Statistics in the SCL and CLS Designs

Statistic	Relative Bias	Relative Variance	Relative Mean Square Error
SCL Means	0.111	0.165	0.275
CLS Means	0.030	0.094	0.125
SCL Correlation Coefficients	0.020	0.094	0.114
CLS Correlation Coefficients	0.021	0.100	0.121
SCL Path Coefficients	0.013	0.098	0.111
CLS Path Coefficients	0.011	0.096	0.107
SCL Multiple Correlation Coefficients	12.6	3.8	16.2
CLS Multiple Correlation Coefficients	7.8	2.5	10.2

designs. The size of the RELMSE for the Multiple Correlation Coefficients implies that no credence could be given to the values obtained.

Table 4.3.5 gives the proportion of times that the 't' ratio falls within certain intervals for each type of statistic. The appropriate number of degrees of freedom is 5 and the theoretically correct proportions are given in Table 4.1.1. The Multiple Correlation Coefficients do not

Table 4.3.5 Proportion of Times that 't' Ratio Falls Within Selected Intervals (SCL and CLS samples)

Statistic	±2.576	±1.960	±1.645
SCL Means	0.976	0.904	0.848
CLS Means	0.968	0.920	0.888
SCL Correlation Coefficients	0.952	0.904	0.840
CLS Correlation Coefficients	0.932	0.856	0.796
SCL Path Coefficients	0.960	0.916	0.880
CLS Path Coefficients	0.960	0.876	0.804
SCL Multiple Correlation Coefficients	0.880	0.856	0.827
CLS Multiple Correlation Coefficients	0.920	0.880	0.867

Table 4.4.1 Average Deft Estimates (STR samples)
Average 'N' was 144.56

Statistics	Empirical	Estimated	Percent error of estimator
Means	0.7419	0.5629	-24.1
Correlation Coefficients	0.7700	0.5120	-33.5
Path Coefficients	0.8532	0.5446	-36.2
Multiple Correlation Coefficients	0.7203	0.6527	-9.4

seem quite so disastrous in this table, but in fact the averaging process has concealed three extreme results. The other statistics seem to be giving a reasonable approximation to a 't' distribution with the case of the Correlation Coefficients in the CLS design being more spread out than the rest.

4.4 Results for the Stratified Samples: STR and WTD

The deft estimates for both these designs are given in Table 4.4.1 and 4.4.2.

For all cases but one the Woodruff estimator is considerably lower than one would wish. When this is converted to a probability statement in Table 4.4.3 the interpretation is clear. With the possible exception of Multiple Correlation Coefficients, the Woodruff estimator is considerably biased. These calculations were carried out on the assumption that the appropriate number of degrees of freedom was the number of samples cases minus the number of strata; this is the way that Frankel calculated degrees of freedom in his study (Frankel, 1971:79). He expressed the situation as 'the hypothesized degrees of freedom are H, the number of strata ..' which, as he was working with only two cases per stratum works out to the same as the usual formula. Suppose however that the quoted hypothesis were correct no matter how many cases there were in each stratum. If this were true, then the probabilities would have to be recalculated on the basis of only three degrees of freedom. This has been done and the results are shown in parenthesis beside the original figures in Table 4.4.3. These latter results are more reasonable than the former, but are still not very encouraging.

Table 4.4.2 Average Deft Estimates (WTD samples)

Average 'N' was 146.08

Statistics	Empirical	Estimated	Percent error of estimator
Means	2.6408	1.0809	-59.1
Correlation Coefficients	1.4477	1.0084	-30.3
Path Coefficients	1.4680	1.0428	-29.0
Multiple Correlation Coefficients	1.2249	1.4317	16.9

An alternative exploration of these poor STR results is to consider the number of cases used to estimate the U-statistics within each stratum. Table 3.2.3 indicates that 103 cases from stratum 1 were used, 34 from stratum 2, and only 13 from stratum 3. When using the Woodruff-Causey program in its stratified mode, separate estimates of all the derivatives are made for each stratum for each relevant variate. There is only one such variate for each of the Means, but there are five for each of the Correlation Coefficients and up to 20 for the Path Coefficients and Multiple Regression Coefficients. It would seem a dubious practice to calculate 20 derivatives from as few as 13, or even 34, cases. One solution to this problem would be to run the program in its population mode, making appropriate corrections to the variance subroutine.

Table 4.4.3 Probability of an Incorrect Statement About the Statistics in the STR and WTD Designs

	Probability of incorrect statement when a two-sided 95% confidence interval is to be used	
Statistic	STR design	WTD design
Means	0.137 (0.095)	0.422 (0.283)
Correlation Coefficients	0.193 (0.125)	0.173 (0.113)
Path Coefficients	0.211 (0.135)	0.164 (0.109)
Multiple Correlation Coefficients	0.076 (0.063)	0.022 (0.034)

Note: Results calculated on 3 degress of freedom are in parentheses.

Table 4.4.4 Bias and Variance Contributions to the Relative Mean Square Error for the Statistics in the STR and WTD Designs

Statistic	Relative Bias	Relative Variance	Relative Mean Square Error
STR Means	0.069	0.001	0.071
WTD Means	0.308	0.003	0.311
STR Correlation Coefficients	0.116	0.003	0.119
WTD Correlation Coefficients	0.102	0.008	0.110
STR Path Coefficients	0.132	0.004	0.135
WTD Path Coefficients	0.088	0.012	0.100
STR Multiple Correlation Coefficients	0.297	0.091	0.388
WTD Multiple Correlation Coefficients	0.510	0.354	0.865

For the WTD sample design there were 50 cases for each stratum. This may well be insufficient for good results. The effect of the weighting process on the Woodruff-Causey program may also be quite negative. However, the evidence is insufficient to make any firm conclusions.

The relative contributions of bias and variance to RELMSE are given in Table 4.4.4. The variance contribution, except for the Multiple Correlation Coefficients, conform to the pattern of the SRS sample, whilst the bias contributions are quite uniformly high. The high variance contribution for the Multiple Correlation Coefficients is an interesting counterpoint to the relative accuracy of the deft estimates.

The proportion of times that the 't' ratio falls within selected limits is given in Table 4.4.5. The STR results here bear out the speculation that the appropriate number of degrees of freedom could well be as low as three. There seems to be no recognizable pattern to the WTD results. Once again the Multiple Correlation Coefficients successfully avoid fitting what little pattern does emerge here.

The poor results for multiple correlation coefficients were not unexpected. The simulation study by Frankel (1971) also produced poor sampling error estimates for all three single-sample techniques under investigation. In a later paper, Kish and Frankel (1974) attribute this poor performance to the problem of using the multiple correlation coefficient with multinomial data (Kish and Frankel, 1974:19 and 35).

Table 4.4.5 Proportion of Times that 't' Ratio Falls Within Selected Intervals (STR and WTD Designs)

Statistic	±2.576	±1.960	±1.645
STR Means	0.952	0.912	0.808
WTD Means	0.696	0.600	0.560
STR Correlation Coefficients	0.912	0.812	0.712
WTD Correlation Coefficients	0.876	0.812	0.732
STR Path Coefficients	0.880	0.764	0.680
WTD Path Coefficients	0.916	0.809	0.729
STR Multiple Correlation Coefficients	0.760	0.733	0.707
WTD Multiple Correlation Coefficients	0.880	0.787	0.747

4.5 Comparison with other Single-Sample Techniques

Ross (1976:46-50) used two other single-sample techniques to estimate sampling errors. For the CLS design he used a Jackknife technique. The results are given in Table 4.5.1. These results should be treated with caution as they are derived from only one example of the CLS design. On comparing the percent errors in deft with those found for the Woodruff technique, the Woodruff technique appears perhaps just a little superior. Turning to the results for the WTD design in Table 4.5.2, the results for both techniques are so poor that comparison is not rewarding.

Table 4.5.1 Results of Application of the Jackknife to One Example of the CLS Design: Deft Estimates

Statistic	Empirical	Estimated	Percent Error of Estimator
Means	2.80	3.09	10.4
Correlation Coefficients	1.53	1.63	6.5
Path Coefficients	1.47	1.53	4.1
Multiple Correlation Coefficients	1.31	1.44	9.9

(after Ross, 1976:47-50)

Table 4.5.2 Results of Application of Balanced Repeated Replication to an Example of the WTD Design: Deft Estimates

Statistic	Empirical	Estimated	Percent Error of Estimator
Means	2.89	4.12	42.6
Correlation Coefficients	1.85	1.66	-10.3
Path Coefficients	1.73	1.63	-5.8
Multiple Correlation Coefficients	2.14	1.20	-43.9

(after Ross, 1976:47-50)

4.6 Summary

The Woodruff-Causey program has been found to give accurate and stable estimates of the statistics in the SRS, SCL and CLS sample design, with the exception of the Multiple Correlation Coefficients in the two clustered designs. This exception is troublesome as educational researchers would usually not have the means of checking that the sampling errors generated by the program had not 'inflated' as they did in this case.

The results for the stratified designs were not so encouraging although the fact that most of the estimators were quite stable leads one to suspect that it may be possible to arrive at some bias correction factor with further work. The poor results all occurred in cases where there was some support for the idea that the samples sizes may have been unreasonably small. This raises the point that this technique is not a way of compensating for inferior sample design. If anything, accurate sampling error estimation for higher-order statistics requires *better* samples than those found adequate to estimate the first-order statistics.

CHAPTER 5

CONCLUSION

In this study an empirical sampling approach has been used to assess the accuracy of an approximation technique for the estimation of sampling errors in several sampling situations commonly used by educational researchers. The investigation was limited to four types of statistics used in correlational and regression studies - the mean, correlation coefficient, path coefficient and the multiple correlation coefficient. When applied to the simple random sample situation and the clustered designs, the technique provided useful estimates for all the statistics except for the multiple correlation coefficient; the problem of sampling error estimation for this statistic has been noted in previous research (Kish and Frankel, 1974:35). The quality of the estimates declined considerably however for the stratified designs; this leads to speculation that the technique might only be reliable in cases where the minimum size of the strata is reasonably high.

Table 5.1 gives some indication of the importance of finding a successful solution, or at least an arsenal of strategies to cope with, the problem of estimating sampling error. Here are displayed the probabilities of an incorrect statement under a 95 per cent confidence interval, which would hold if the design factor were to be ignored: note that in such a case the researcher has assumed that simple random sampling gives an adequate approximation to the sample design which was employed and hence, that all these probabilities are not too far from 0.050. Patently, any inferences made under these assumptions will be entirely untenable for the SCL, CLS and WTD designs, whilst for the SRS and STR designs, the simple random sample assumption has led to rather conservative confidence intervals.

The Woodruff-Causey program has been shown here to provide a significant improvement on this performance for cases where the effective sample size is not too small. The program can give an estimate of the sampling error for any statistic which can be expressed as a Fortran subroutine; the user need only supply this subroutine and, depending on the circumstances, a subroutine to estimate variance and a few data-manipulation subroutines. For more standard situations several less flexible but less demanding programs (which were mentioned in Section 2.6) are now available.

Table 5.1 Probability of Incorrect Statements When the Design Factor is Ignored

Design	Means	Probability of incorrect statement when a two-sided 95% confidence interval is used: Correlation Coefficients	Path Coefficients	Multiple Correlation Coefficients
SRS	0.05	0.03	0.05	0.01
STR	0.03	0.04	0.04	0.04
SCL	0.21	0.09	0.09	0.07
CLS	0.48	0.20	0.18	0.13
WTD	0.50	0.29	0.26	0.36

Note: The first row is taken from Table 4.2.3 of this study; the rest are taken from Ross (1976:39-45).

In common with these other programs using the Taylor's series approximation, the Woodruff-Causey program enjoys the advantages of a relatively high computational speed and transparency of assumptions. However, it also handsomely repays the demands it makes on the skills of the researcher with the marked flexibility it displays in handling diverse sampling situations for estimating the sampling errors of almost any statistic imaginable and in its adaptability to quite small computer installations.

The results have indicated the need for further evaluation of the technique in situations where larger number of cases are involved especially for stratified and weighted sample designs.

Although the results of this study are only empirical estimates based on particular sampling schemes and for particular statistics, the pattern of results is most probably applicable to a wide range of studies undertaken by educational researchers. Considering the broad range of possible sample designs and statistical analyses which are available to the educational research worker, it would seem doubtful that a comprehensive theoretical solution to the problem of sampling error will ever become available. However, the problem is with us now, and approximation methods such as the Woodruff-Causey program, if used cautiously, have been shown in this study and elsewhere, to give stable estimates of the often large sampling errors present in educational research data.

REFERENCES

Bachman, J.G., R.L. Kahn, M.T. Medvick, T.N. Davidson and L.D. Johnston.
1967 Youth in Transition, volume 1: Blueprint for a longitudinal study of Adolescent Boys. Ann Arbor: Survey Research Centre, University of Michigan.

Bean, J.A.
1975 Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution. US Department of Health, Education and Welfare publication No. HRS-75-1339.

Bobko, P. and A. Rieck.
1980 Large sample estimators for standard errors of functions of correlation coefficients. Applied Psychological Measurement, 4, 385-398.

Causey, B.D.
1976 Computer program for the computation of sampling errors of complicated estimates: technical documentation. Mimeographed paper, Washington, Statistical Research Division, Bureau of the Census.

Broom, L., P. Duncan-Jones and P. McDonnell.
1977 Investigating Social Mobility. Canberra: A.N.U. Sociology Department Monograph No. 1.

Cochran, W.G.
1963 Sampling Techniques. 2nd edition. New York: John Wiley.

Comber, L.C. and J.P. Keeves.
1973 Science Education in Nineteen Countries. Stockholm: Almqvist and Wiksell/New York: Wiley-Halsted.

Deming, W.E.
1956 On simplifications of sampling design through replication with equal probabilities and without stages. Journal of the American Statistical Association, 51, 24-53.

Deming, W.E.
1960 Sample Design in Business Research. New York: John Wiley.

Duncan, O.D.
1975 Introduction to structural equation models. New York: Academic Press.

Finch, R.H. Jnr.
1978 Estimating sampling variability in a multi-survey environment. Proceedings of the Survey Research Section of the American Statistical Association, 392-394.

Finifter, B.M.
1972 The generation of confidence: evaluating research findings by random subsample replication. In H.L. Costner (Ed.) Sociological Methodology, 1972. San Francisco: Jossey-Bass.

Frankel, M.R.
1971 Inference from Survey Samples. Ann Arbor: Institute for Social Research, University of Michigan.

Guilford, J.P. and B. Fruchter.
1973 Fundamental Statistics in psychology and education. Tokyo: McGraw-Hill, Fifth Edition.

Hansen, M.H., W.N. Hurwitz and W.G. Madow.
1953 Sample Survey Methods and Theory, Vols. I and II. New York: Wiley.

Henrici, P.
1964 Elements of Numerical Analysis. New York: Wiley.

Hidiroglou, M.A., W.A. Fuller and R.D. Hickman.
1975 SUPER CARP. Ames, Iowa: Survey Section, Iowa State University.

Holt, M.M.
1977 SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data. Unpublished report, Research Triangle Institute, Research Triangle Park, North Carolina.

Kaplan, B., I. Francis and J. Sedransk.
1979 A comparison of methods and programs for computing variances of estimates from complex survey samples. Proceedings of the Survey Research Section of the American Statistical Association, 97-100.

Keeves, J.P.
1966 Students attitudes concerning mathematics. Unpublished MEd thesis, University of Melbourne.

Keeves, J.P.
1971 The home, the school and educational achievement. Unpublished PhD. thesis, Australian National University.

Kendall, M.G. and A. Stuart.
1963 The Advanced Theory of Statistics. Volume I. London: Griffin.

Kerlinger, F.N. and E.J. Pedhazur.
1973 Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston.

Kish, L.
1965 Survey Sampling. New York: John Wiley.

Kish, L.
1969 Design and estimation for subclass comparisons and analytic statistics. In N.L. Johnston and H. Smith (eds.) New Developments in survey sampling. New York: Wiley.

Kish, L. and M.R. Frankel.
1970 Balanced Repeated Replications for standard errors. Journal of the American Statistical Association, 65, 1071-1094.

Kish, L. and M.R. Frankel.
1974 Inference from Complex Samples. Journal of the Royal Statistical Society, Series B, 36, 1-37.

Krewski, D. and J.N.K. Rao.
1978 Inference from stratified samples 1: Large sample properties of the linearization, Jackknife and Balanced Repeated Replication Methods. Carleton Mathematical Series No. 155, Carleton University, Ottowa.

Krewski, D. and J.N.K. Rao.
1979 Small sample properties of the linearization, jackknife and balanced half-sample methods for ratio estimation in stratified samples. Paper presented at the Annual Meeting of the American Statistical Association.

Mahalanobis, P.C.
1944 On Large Scale Sample Surveys. Philosophical Transactions of the Royal Society. B231. 329-451.

Mahalanobis, P.C.
1946 Recent Experiments in Statistical Sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, 109, 325-451.

Marks, E.S.
1947 Sampling in the revision of the Stanford-Binet Scale. Psychological Bulletin, 44, 413-444.

Maurer, K., G. Jones and E. Bryant.
1978 Relative Computational efficiency of the linearized and balanced repeated replication procedures for computing sampling variances. Proceedings of the Survey Research Section of the American Statistical Association, 388-391.

McCarthy, P.J.
1966 Replication: an approach to the analysis of data from complex surveys. National Center for Health Statistics, Series 2. No. 14.

McNemar, Q.
1969 Psychological statistics. (4th ed.) New York: Wiley.

Mellor, R.W.
1973 Subsample replication variance estimation. PhD thesis, Harvard University.

Miller, R.G. Jnr.
1968 Jackknifing Variances. Annals of Mathematical Statistics, 39, 567-582.

Mosteller, F. and J.W. Tukey.
1968 Data analysis including statistics. In G. Lindzey and E. Aronson (Eds). The Handbook of social psychology. 2nd edition. Reading, Massachusetts: Addison-Wesley.

Mosteller, F. and J.W. Tukey.
1977 Data analysis and regression: a second course in statistics. Reading, Massachusetts: Addison-Wesley.

Peaker, G.F.
1953 A sampling design used by the Ministry of Education. Journal of the Royal Statistical Society, 116, 140-165.

Peaker, G.F.
1967 Sampling. In T. Husen (Ed.) International study of achievement in mathematics, volume 1. Stockholm: Almqvist and Wiksell.

Peaker, G.F.
1975 An empirical study of education in twenty-one countries: a technical report. Stockholm: Almqvist and Wiksell.

Pearson, E.S. and J. Wishart.
1947 'Students' Collected Papers. Biometrika Office, University College, London.

Quenouille, M.J.
1956 Notes on bias in estimation. Biometrika, 43, 353-360.

Ross, K.N.
1976 Searching for uncertainty. An empirical investigation of sampling errors in educational survey research. Hawthorn: ACER. Occasional Paper No. 9.

Shah, B.V.
1974 STDERR: Standard Errors Program for Sample Survey Data. Research Triangle Institute, Research Triangle Park, North Carolina.

Shah, B.V.
1978 Variance estimates for complex statistics from multistage sample surveys. In Namboordiri, N.K. (Ed.) Survey sampling and measurement. New York: Academic Press.

Sukhatme, P.V.
1954 Sampling Theory of Surveys with Applications. New Delhi: The Indian Society of Agricultural Statistics and The Iowa State College Press.

Tepping, B.J.
1968 The estimation of variance in complex surveys. Proceedings of the Social Statistics Section of the American Statistical Association, 11-18.

Tukey, J.W.
1954 Unsolved Problems of experimental statistics. Journal of the American Statistical Association, 49, 706-731.

Tukey, J.W.
1958 Bias and confidence in not-quite large samples: abstract. Annals of Mathematical Statistics, 29, 614.

Verma, V. and M.C. Pearce.
1978 User's Manual for CLUSTERS. London: World Fertility Survey.

Verma, V., C. Scott and C. O'Muircheartaigh.
1980 Sample designs and sampling errors for the World Fertility Survey. Journal of the Royal Statistical Society, Series A, 143, 431-473.

Wilson, M.
1981 Adventures in Uncertainty: An empirical investigation of the use of a Taylor's series approximation for the assessment of sampling errors in educational research. University of Melbourne. Unpublished MEd Thesis.

Woodruff, R.A.
1971 Simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association, 66, 411-414.

Woodruff, R.A. and B.D. Causey.
1976 Computerized method for approximating the variance of a complicated estimate. Journal of the American Statistical Association, 71, 315-321.

Occasional Paper 17

This study investigates the accuracy of the Woodruff-Causey technique for estimating sampling errors for complex statistics. The technique may be applied when data are collected by using multi-stage clustered samples. The technique was chosen for study because of its relevance to the correct use of multivariate analyses in educational survey research. A guide to the use of the technique and to writing the relevant Fortran sub-routines is included in microfiche appendixes.

The study also includes a review of the literature in the field.

AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH

ISBN 0 85563 469 3

APPENDIX J

NUMERICAL DIFFERENTIATION

If f is the statistic under investigation using the Woodruff-Causey program, then one of the important steps is the evaluation by numerical methods of the partial derivatives,

$$\frac{\partial f(V_1, V_2 \ldots V_r)}{\partial V_i} \quad i=1, \ldots r$$

at the expected values of the sums of the variates V_i. In fact, these expected values can be evaluated only by using the actual sample values $v_1, v_2 \ldots v_r$. The expression used to find the partial derivatives is

$$f'_n = \frac{1}{2h}\left(f(v_1, v_2 .. v_i+h .. v_r) - f(v_1, v_2 .. v_i-h .. v_r)\right)$$

for i=1, 2.. r (1)

This is a straight-forward application of the usual definition of a partial derivative:

$$\frac{\partial f}{\partial v_i} = \lim_{h \to o} \frac{1}{2h}\left(f(v_1, v_2 .. v_i+h .. v_r) - f(v_1, v_2 .. v_i-h .. v_r)\right) \quad (2)$$

The only difficulty in applying the approximation (1) is in choosing a suitable value for h. This is found by considering the possible errors involved in the approximation.

It may be shown (Henrici, 1964:236) that the error involved in the approximation is

$\frac{1}{6} h^2 f'''(\xi)$ (where f' is used to denote the theoretical first derivative, etc)

where $v_i - h < \xi < v_i + h$. In addition, one must consider the machine-error in the calculation of $v_i + h$ and $v_i - h$, which is approximately bounded by $2 \left| v_i f' \right| P$, where $P = 10^{-M}$ and M is the number of significant figures used by the machine. When these values are used to find the two estimates of F, a further error bounded approximately by

$$f(v_1, v_2 \ldots v_r)\, P$$

is also involved (Woodruff and Causey, 1976 : 321). For a non-zero partial derivative, the relative error is then bounded approximately by

$$\frac{1}{6} \left| \frac{f'''}{f'} \right| h^2 + \left(\tfrac{1}{2} \left| \frac{f}{f'} \right| + \left| v_i \right| \right) \frac{P}{h} \qquad (3)$$

Obviously, as h gets smaller the first term will decrease but, since P is fixed, the second term will increase. Thus the strategy is to choose h as small as possible without

$$\frac{1}{h} \left(\tfrac{1}{2} \left| \frac{f}{f'} \right| + \left| v_i \right| \right)$$

becoming to large. The program uses an iterative procedure to find an appropriate h according to the steps outlined above and, of course uses f_n' to approximate f'. The only extra problem occurs where f' is either zero or very near to it: in this case the iterative procedure is very slow with the possibility that h would need to be very large before

$$\frac{1}{h} \left(\tfrac{1}{2} \left| \frac{f}{f'} \right| \right)$$

becomes small. To circumvent this problem, f' is set to zero when h exceeds $\frac{|v_i|}{1000}$.

APPENDIX K

A USER'S GUIDE TO THE WOODRUFF-CAUSEY PROGRAM FOR THE COMPUTATION OF THE SAMPLING ERROR OF COMPLICATED ESTIMATES

The following user's guide has been written in an attempt to 'soften' the rather technical documentation which accompanied the program (Causey, 1976). Potential users must be warned however, that only those with more than a beginners knowledge of Fortran should attempt to use the program. Although the program demands some writing of Fortran subroutines, the user will find that such efforts are well-rewarded; for the program exhibits great flexibility not only in the type of sampling problem it can handle, but also in the procedures it uses to solve the problem. Furthermore, in an environment where particular sampling situations were the norm, it would not be difficult to set-up the program to handle such standard situations without the need for subroutine writing. The following is based on the technical documentation which accompanied the program; any errors are, of course, the responsibility of the present author.

K1 A Worked Example

The example which follows was chosen as one that would indicate the steps necessary to use the Woodruff-Causey program, and yet be simple enough to provide an introduction to the technique. Consequently issues such as weighting, the use of temporary storage space, and the use of a user-written variance subroutine are left to the formal description of the program in Sections K4 and K5.

K1.1 The problem

Suppose a researcher wishes to investigate how ambitious are young secondary students, and how this might relate to their ethnic origins. Some data is collected consisting of the students' opinions as to their later occupations, the present occupations of their fathers, and the language spoken in the home. The data are coded according to Table K1.1. The scale of occupational prestige is the six-point ANU scale (Broom et al., 1977:112). In order to make the occupational prestige scale amenable to a product-moment correlation investigation, the occupational categories are transformed into an approximately interval level scaled score as in Table K1.2.

Table K1.1 Format of Input Data for 'A Worked Example'

Variable	Columns	Format	Comments
ID	1-3	I3	Identification number
FOCCUP	4	I1	Six-point scale of occupational prestige
EXPOCC	5	I1	Six-point scale of occupational prestige
ENSPKHME	6	I1	English spoken in parental home = 1
			A language other than english spoken in parental home = 0
			Missing data = 2

Table K1.2 The Six-Point ANU Scale of Occupational Prestige

Occupational grouping	Rank	Weighted Social status score
Professional	1	662
Managerial	2	611
White Collar	3	508
Skilled Manual	4	485
Semi-skilled Manual	5	421
Unskilled Manual	6	418
Missing Data	7	-

An index of ambition is formed in the following way. If o_i and f_i are a student's scores on EXPOCC and FOCCUP respectively, then define a measure of ambition as

$$a_i = o_i - f_i \qquad (1)$$

Then if e_i is the student's score on ENSPHIME, find the product moment correlation a:

$$r_{ae} = \frac{n\sum_{i=1}^{n} a_i e_i - n\bar{a}\bar{e}}{\sqrt{\left(n\sum_{i=1}^{n} a_i^2 - n\bar{a}^2\right)\left(n\sum_{i=1}^{n} e_i^2 - n\bar{e}^2\right)}} \qquad (2)$$

where n is the number of cases in the sample,

$\bar{a}$ is the average of the a_i

and $\bar{e}$ is the average of the e_i.

For this data, the mean of the a_i is found to be 56.24 and the correlation between the a_i and the e_i is -0.1054.

The sampling scheme used to collect the data was a simple random sample of 600 cases with replacement, so the usual

estimator of sampling error can be used, that is,

$\frac{s}{\sqrt{n}}$ for the average of the a_i

$\frac{1}{\sqrt{n}}$ for the correlation coefficient

where s is the standard deviation of the a_i within the sample.

For the data, these take the values 3.95527 and 0.0436021 respectively, when all cases with any missing values are deleted. Thus the correlation between language spoken at home and this index of ambition would seem to be weak but non-random, at least at a 95 per cent confidence level. However, the researcher is, quite understandably, concerned with the use of error estimates which involve assumptions of normality when one of the variables is clearly not normally distributed. The Woodruff-Causey program can be used to clarify the situation.

K1.2 Solving the problem

The program is capable of solving this problem in a number of superficially different ways. The actual manipulations of the data will be the same in each possible arrangement, but the ways in which the control information and the data are fed to the program will differ markedly. The program will always need the following types of information in some way.

- the number of strata in the sample, and certain information about each stratum.

2 The number of functions to be investigated, and a 'formula' in the form of a Fortran segment for each.

3 The number of variates involved in the investigation, an indicator of which variates are relevant to which function, and the actual data on which the investigation is to be made.

4 Auxiliary information such as extra output, tempora- storage, etc.

One particular way of solving the problem is descri[illegible]
A solution starts with the 'Problem Card'.

The program interprets the information contained on that card in the following way.

1 A value of 1 in columns 1 to 4.
This means there is only one stratum to consider.

2 A value of 2 in columns 5 to 8.
This means there are two functions whose sampling errors are under consideration.

3 A value of 5 in columns 9 to 12.
This means there are five variates involved in the problem.

4 A value of 1 in column 13.
This means that the program must look to subroutine NSTRAT for stratum information.

5 A value of 1 in column 16.
This means that the program must look to subroutine NINPUT for the data.

6 A value of 1 in column 20.
This means that certain information concerning each derivative is to be printed out.

7 A value of 1 in column 35.
This means that not all the variates are involved in all the functions.

One important point must be made, the Fortran function F must calculate the functions under investigation using not the individual values of the *variables*, but the sums over the entire population of each of the *variates*. The distinction is important, the variables are the measures which are under investigation whilst the variates are the variables plus certain transformation of the variables which will be needed in the calculation of the function. Thus, in this case the variables are as given in Table K1.1, but the variates, which are the values to be read into the program, are quite different. In order to calculate the two functions, the following five sums are needed:

$$\Sigma e_i, \quad \Sigma e_i^2, \quad \Sigma a_i, \quad \Sigma a_i^2, \quad \Sigma a_i e_i$$

Hence, for each case the input data must be

$$e_i, \ e_i^2, \ a_i, \ a_i^2, \ a_i e_i$$

In general there will be more variates than variables. The composition of this list of variates is not unique. For instance, in this case it would have been quite possible to write a Fortran function which calculated the mean and correlation coefficient in terms of the following set of

variates:

$$e_i,\ e_i^2,\ o_i,\ f_i,\ a_i^2,\ a_i e_i$$

Having described how the data is to be treated on this problem card, the user must then write several Fortran segments; always a Main Program and a function, F, and, if the user has so indicated on the Problem Card, any others which are necessary.

The Main Program. This is used to start the program executing, to reserve dimensional array space and common

```
C     PROGRAM DINKY
      DIMENSION A(3821),N(7)                    } Reserve space for other
      DOUBLE PRECISION D(15)                    } subroutines (See K3.2.1)
      DIMENSION IX(3),W(600,5)
      COMMON/COMMN/W,K3                         Common Block for MINPUT
      COMMON/COMMN/NSEL                         Common Block for MSTRAT
      K3=0
      I=1
      WRITE(6,2)
 2    FORMAT(' SAMPLE VARIANCE FOR DINKY EXAMPLE')
      DO 50 K=1,600
      READ(5,10) (IX(II),II=1,3)
 10   FORMAT(3X,3I1)
      IF(IX(1).EQ.7) GO TO 50                   } Likewise deletion of
      IF(IX(2).EQ.7) GO TO 50                   } missing data
      IF(IX(3).EQ.9) GO TO 50                   }
      DO 20 J=1,2
      IF(IX(J).EQ.1) IX(J)=662                  }
      IF(IX(J).EQ.2) IX(J)=611                  } Manipulation of variables
      IF(IX(J).EQ.3) IX(J)=508                  } IXCOLP and IXSCOL to make
      IF(IX(J).EQ.4) IX(J)=405                  } them interval level
      IF(IX(J).EQ.5) IX(J)=421                  }
      IF(IX(J).EQ.6) IX(J)=418                  }
 20   CONTINUE
      W(I,1)=FLOAT(IX(3))                       }
      W(I,2)=W(I,1)*W(I,1)                      } Formation of the variates;
      W(I,3)=FLOAT(IX(2))*FLOAT(IX(1))          } they are stored in
      W(I,4)=W(I,3)*W(I,3)                      } COMMON
      W(I,5)=W(I,1)*W(I,3)                      }
      NSEL=I                                    ... Keeping track of number
      I=I+1                                     of cases not deleted
 50   CONTINUE
      WRITE(6,60) NSEL
 60   FORMAT(' NSEL IS ',I3)
      CALL PREPARE(A,3821,N,7,D,15)             ... Hand control to PREPARE
      STOP
      END
```

Figure K1.1 Program DINKY

blocks for later subroutines, to carry out any calculations which need be performed only once, and to call the first of the supplied subroutines, PREPAR.

Note that data cases with missing values are eliminated entirely from the calculations. This is necessary as at a later stage a linear combination of the variates is to be calculated for each case (the result of this operation is called a 'U-statistic').

Subroutine WINPUT. This was requested by the fifth entry on the Problem Card. Its function is to supply the five variates, one case at a time, to PREPAR. In this example the variates have been placed in the Common block COMMW by the Main Program, so all that this subroutine need do is transfer the cases in the correct order back to the calling subroutine through the argument W.

```
      SUBROUTINE WINPUT(W,N)
      DIMENSION W1(600,5),W(N)
      COMMON/COMMW/W1,K3
      K3=K3+1
      DO 10 J=1,5
      W(J)=W1(K3,J)
10    CONTINUE
      RETURN
      END
```

Figure K1.2 Subroutine WINPUT

Subroutine NSTRAT. This was requested by the fourth entry on the Problem Card. Its function is to supply four pieces of information for each stratum to the calling subroutine. The information needed is N, the number of cases in the stratum, 'FR', the sampling fraction, 'NT', the total number

of cases (which should be set to zero if FR is supplied) and a quantity named IV which tells the program how to estimate the variance. In this case, there is only one stratum, the number of cases has been placed in Common block COMMN by the main program, the sampling fraction is 0.0 since we have sampling with replacement, and the value of IV is 0 which indicates that the program is to use its default simple random sampling variance formula.

```
SUBROUTINE NSTRAT(I,N,FR,NT,IV)
COMMON/COMMN/NSEL
N=NSEL
FR=0.0
NT=0
IV=0
RETURN
END
```

Figure K1.3 Subroutine NSTRAT

The Function F. This must always be supplied by the user. Its purpose is to calculate the functions under investigation using the variate sums (which in this case are contained in the first argument, T) in the order in which they were supplied from WINPUT. The number of the function is supplied by a Common block called COMMF which is defined within subroutine GENVAR. This function is to be calculated using double precision wherever possible.

```
      DOUBLE PRECISION FUNCTION F(T,NT,S,NS,NS,R,NR)
      DOUBLE PRECISION T(NT),S(NS,NS),R(NR)
      DOUBLE PRECISION FUN(2),V(2)
      COMMON/COMMF/L
      COMMON/COMMN/N
      FUN(1)=T(3)/DFLOAT(N)
      IF(L.EQ.1) GO TO 99
      V(1)=DFLOAT(N)*T(2)-T(1)*T(1)
      V(2)=DFLOAT(N)*T(4)-T(3)*T(3)
      FUN(2)=(DFLOAT(N)*T(5)-T(1)*T(3))/DSQRT(V(1)*V(2))
   99 F=FUN(L)
      RETURN
      END
```

Figure K1.4 The Function 'F'

The Subroutine NSUBFV. This subroutine is requested by the last entry on the Problem Card. Its purpose is to inform the program of which variates are involved in the calculation of each function. In this case, function 1, the mean, involves only variate 3, whereas function 2, the correlation coefficient involves all of the variates. The subroutine has three arguments; the first is the number of the funciton, L, the second is the number of the variate, J, and the third is a value to be supplied, which equals two if the variate J is involved in the calculation of the function L, and equals one if it is not involved.

```
SUBROUTINE NSUBFV(L,J,IFV)
IFV=1
IF(L.EQ.2) IFV=2
IF(J.EQ.3) IFV=2
RETURN
END
```

Figure K1.5 Subroutine NSUBFV

The Dummy Subroutines. Although they are not involved in this example, the Fortran compiler at many installations will require the following two dummy subroutines.

```
SUBROUTINE VARNCE (T,M,I,S,N,F,R,Q,INV)
RETURN
END
SUBROUTINE NSUBHV (I,J,...V)
RETURN
END
```

The first would be needed if the user wished to supply a different variance formula from the simple random sampling formula (which is a default). The second performs a similar function for strata as NSUBFV performs for variates.

The Supplied Subroutines. The 'program' as supplied consists of the three subroutines PREPAR, GENVAR, and SWITCH. These need not be altered.

The Printed Output. The printed output from this example is given in Figure K1.6.

```
SAMPLE VARIANCE FOR DINKY EXAMPLE
NSEL = 526
DIMENSION LIMITS ARE   2648     7     7

DERIVATIVES FOR FUNCTION      1
       3  3  0.17011407E-02  0.75710065D-03


FUNCTION       1  0.55237840E+02  0.15643397E+02  0.39552355E+01  0.70337739E-01

DERIVATIVES FOR FUNCTION      2
       1  3 -0.67374047E-02  0.16150077D-04
       2  3  0.13346771E-02  0.16150079D-04
       3  3 -0.71677124E-04  0.75910960D-03
       4  3  0.12200431E-07  0.25335101D+00
       5  3  0.78536113E-04  0.86225701D-03


FUNCTION       2 -0.10541481E+00  0.17737219E-02  0.42137400E-01  0.39974836E+00
```

Figure K1.6 The Output from Example DINKY

'NSEL' is the number of cases left after list-wise deletion.

The 'DIMENSION LIMITS' are the required minimum dimension sizes needed for the arrays A, N and D respectively, of the Main Program. The derivative information for each function consists of the number of the variate involved, the number of iterations needed to achieve stability, the value of the derivative, and the final increment.

This information given for each function is

1 the number of the function

2 the estimated value of the function

3 the estimated variance of the function

4 the estimated standard deviation of the function

5 the coefficient of variation (i.e. (4) divided by (2))

The results indicate that the standard error for the mean given by the usual formula was as accurate as one could ever expect, but then its accuracy was not in question. For the correlation coefficient, however, the program has revealed a 3.5 per cent inaccuracy in the estimate of error given by the usual formula. It is instructive moreover that this inaccuracy does not alter the earlier verdict regarding the correlation coefficient.

K2 A Formal Description of the Situation

K2.1 Non-Stratified Case

For a certain population, consider V variates with totals $Y_1, Y_2 \ldots Y_v$ over the population which are combined in some

way in a function F of all V variate totals. The goal is to estimate Var (F).

Suppose a sample is now drawn from the population and in that sample X_{ij} is the value of variate i for case j. If, furthermore, P_j is the probability (prior to inclusion) of j in the sample, let

$$W_{ih} = \frac{X_{ij}}{P_j}$$

and then

$$\hat{Y}_i = \sum_{j=1}^{n} W_{ij},$$

when n is the total number of cases in the sample, is an estimator of Y_i, the population sum.

This allows one to estimate

$$F = F(Y_1, Y_2 \ldots Y_v) \text{ with}$$

$$\hat{F} = F(\hat{Y}_1, \hat{Y}_2 \ldots \hat{Y}_v).$$

The first-order Taylor Expansion of $\hat{F}$ about F is then

$$F_2 = F + \sum_{i=1}^{v} D_i(\hat{Y}_i - Y)$$

where D_i is the derivative

$$\frac{\partial F}{\partial Y_i} \quad \text{evaluated at } (Y_1 \; Y_2 \ldots Y_v)$$

Var (F) is approximated by Var ($\hat{F}$) which is then approximated by Var (F_2).

The variance of F given by

$$\text{Var}\left(\sum_{i=1}^{v} D_i \hat{Y}_i\right) = \text{Var}\left(\sum_{i=1}^{v} D_i \sum_{j=1}^{n} W_{ij}\right)$$

$$= \text{Var}\left(\sum_{j=1}^{n}\sum_{i=1}^{v} D_i W_{ij}\right)$$

$$= \text{Var}\left(\sum_{j=1}^{n} U_j\right)$$

$$\text{where } U_j = \sum_{i=1}^{v} D_i W_{ij} = \frac{\sum_{i=1}^{v} D_i X_{ij}}{P_j}$$

These U_j's are called 'U-statistics'.

The variance of this total is found in the same way as one would find the variance of each $\hat{Y}_i$ considering the original sampling scheme. Where the sampling scheme was simple random sampling the appropriate variance estimator is

$$\text{Var}\left(\sum_{j=1}^{n} U_j\right) = \frac{(1 - f)\sum_{j=1}^{n} (u_j - \bar{u})^2}{n - 1}$$

where f is the sampling fraction

and $\bar{u}$ is the mean 'U-statistic'.

This formula may be chosen as a default in the program. The calculations may be performed using the weighted W's or the unweighted X's.

K2.2 Independent Strata

If the sampling within each stratum is independent of the rest, the above process may be repeated *within* each stratum. Let H be the number of strata. Use the subscript 'h' to denote that a certain variate or statistic pertains only to stratum 'h', and the above formulae may be rewritten to describe the application of the process to a stratified situation. The variance estimation formula for simple random sampling within each stratum becomes

$$\mathrm{Var}\left(\sum_{j=1}^{n_h} u_{jh}\right) = \frac{(1-f_n)}{} \cdot \frac{\sum_{j=1}^{n_h} (u_{jh} - \bar{u}_h)^2}{(n_h - 1)}$$

K3 Formal Description of the Use of the Program

The required inputs to the program are of three types: first a 'Problem Card' tells the program the type of problem under consideration and the way in which it is to be handled; second, a main program and a series of subroutines must be included with the source deck; third, the data must be provided in the various ways specified in the Problem Card.

K3.1 The Problem Card

The information concerning the type of problem to be investigated and the way that the user wishes it to be handled is supplied to the program in terms of numbers punched on the Problem Card. This card should be left blank unless otherwise indicated. There are seventeen such numbers and the user will be referred to the information immediately below by reference to the names 'PC1', 'PC2' ... 'PC17' which indicate their

sequence on the Problem Card.

K3.1.1 A Column-by-Column Commentary.

PC1 Title: Number of strata Columns: 1-4 Format: I4

Comments: This indicates the number, H, of independent strata. If there are no strata, then H = 1.

PC2 Title: Number of functions Columns: 5-8 Format: I4

Comments: This indicates the number, G, of functions for which variance estimates are desired.

PC3 Title: Number of variates Columns: 9-12 Format: I4

Comments: This indicates the number, V, of variates which are input as data to calculate the value of the functions under consideration. Note that the number of variates will generally be greater than the number of 'variables' involved in the functions under consideration.

PC4 Title: Input Mode for Stratum Descriptors

Column: 13 Format: I1

Comments: For each stratum, the user must supply four quantities. This entry tells the program where to look for them.

If they are to be read from cards, place a '0' in this column.

If they are to be read from an unformatted binary file, place a '1' in this column.

If they are to be provided from a user-written external subroutine NSTRAT, place a '2' in this column.

Details of the inputs to be provided by each of these modes are contained in Section K3.1.2 - The Stratum Descript

PC5 Title: Stratum Descriptors Input File

Columns: 17-18 Format: I2

Comments: If the value of PC4 is '1' then the user must here indicate the internal unit number of the unformatted file on which resides the Stratum Descriptor information..

If the value of PC4 is not '1' then these columns are to be left blank.

Of course the user must ensure that this internal unit number is not used in any other part of the program.

PC6 Title: Input Mode for Variate Values

Columns: 16 Format: I1

Comments: For each sample unit in each stratum the user must provide the V variable values. This column tells the program where to find them.

If they are to read from cards, place a '0' in this column.

If they are to be read from an unformatted binary file, place a '1' in this column.

If they are to be provided from a user-written external subroutine WINPUT, place a '2' in this column.

Details of the inputs to be provided by each of these modes are contained in Section K3.1.2 - Variate Values.

PC7 Title: Variate Values Input File or Number of Format Cards Columns: 17-18 Format: I2

Comments: If the value of PC6 is '0' then the user must here indicate the number of cards needed to give the format of the card or cards from which the V values are to be read for each sample unit in each stratum.
If the value of PC6 is '1' then the user must here indicate the internal unit number of the unformatted binary file on which resides the Variate Values information.
If the value of PC6 is not '0' or '1' then these columns are to be left blank.
Of course the user must ensure that this internal unit number is not used in any other part of the program.

PC8 Title: U-Statistics Columns: 19 Format: I1

Comments: If the user desires a printout of the U-statistics for each case, a '1' is placed in this column.
Note that these are printed in the format

10(1X, E11.6)

Thus the output file or printer must be capable of receiving lines of 120 characters.
If the U-statistics are not needed, leave this column blank.

PC9 Title: Derivatives Columns: 20 Format: I1.

Comments: If the user desires information concerning the derivities of each function with respect to the appropriate variates, a '1' is placed in this column. The information consists of:

1 the number of the variate involved in the differentiation;

2 the number of iterations involved in the numerical procedure (** denotes the maximum of 8);

3 the value of the derivative;

4 the final increment used in the numerical procedure.

This information is provided in 45 characters across each line.

If this information is not needed, leave this column blank.

PC10 , PC11, PC12, PC13, PC14

Collective Title: Temporary Storage Columns: 21-33

Comments: These columns may be left blank unless the amount of space needed for storage of dimensioned arrays, which is discussed in Section K3.2, is beyond the capacity of the particular machine being used. If this is the case, consult Section K4.

PC15 Title: Irrelevant Strata Column: 34 Format: I1

Comments: If all variates are defined for all strata, leave this column blank.

If for some strata, certain variates are not defined, place a '1' in this column. The user must then provide a user-written external function NSUBHV which informs the program of the appropriate pairs of variates and strata. This is described in Section K3.2.3.

Note that an alternative strategy is to supply a value of zero for the variate in the relevant strata. This will result in a loss of efficiency

and may also lead to execution errors.

PC16 Title: Irrelevant Variates Columns: 35 Format: I1

Comments: If, apart from the considerations of Irrelevant Strata (PC15), all functions are defined using stratum totals of *all* of the variates, leave this column blank.

If all functions are defined using population totals of *all* of the variates, place a '2' in this column.

If some variates are not involved in the calculation of some functions, regardless of whether stratum totals or population totals are used, place a '1' in this column.

The user must then provide a user-written external subroutine NSUBFV which informs the program of the appropriate action to be taken with respect to each pair of stratum and variate. This is described in Section K3.2.4.

Note that an alternative strategy of allowing the program to calculate derivatives of irrelevant variates will result in a loss of efficiency and may also lead to execution errors.

The user must be careful at this point to ensure

1. that population sums or stratum sums are used in the function F (see Section K3.2.2) to correspond with what the user has indicated here,

2 that the conditions indicated here do not conflict with those indicated in subroutine NSUBHV (see Section K3.2.6).

PC17 Title: Preliminary Run Columns: 36 Format: I1

Comments: If the user wishes to obtain lower bounds for the dimensioned arrays discussed in Section K3.2.1 without running the entire program, place a '1' in this column.

This step will often not be necessary as all that is needed is to exceed these limits and still remain within the available storage space.

K3.1.2 The Stratum Descriptors. For each stratum the following information must be provided

1 The stratum sample size, N, in integer format.

2 The stratum sampling fraction, FR in floating point format. This is the ratio of stratum sample size to total stratum size.

3 The total stratum size, NT, in integer format. This value is used only to compute the sampling fraction. If NT is set to zero, then FR will be used as the sampling fraction.

4 A value, IV, which informs the program of the mode of variance computation to be employed.

If variance computation is to be done, for the stratum under consideration, using the *internal* simple random sampling variance formula applied to *externally-weighted* data, then set IV equal to 0.

If the same procedure is to be followed, but the data are to be *internally-weighted*, then set IV is equal to 1.
If the variance computation is to be done, for the stratum under consideration, using a user-written external subroutine VARNCE applied to *externally-weighted* data, then set IV is equal to 2.
If the same procedure is to be followed, but the data are to be *internally-weighted*, then set IV is equal to 3.

Where no strata are involved, that is, where only population values are under consideration, this information is to be provided as though for stratum one.

Note that if IV indicates *internal-weighting*, all calculations will be in terms of variate values divided by the probability of selection within the relevant stratum. If IV indicates *external-weighting* the user may pre-weight all data before they are input into the program, or, if using an external subroutine VARNCE, the user may input unweighted data, and make appropriate adjustments within the function and variance subroutines.

The program finds this information in the way indicated by the value of PC4. If PC 4 equals zero, the information is to be read from cards. The first card must give, in columns 1 to 72 and in format (12A6), the format of the H cards, one per stratum, providing for each of the H strata the four quantities. The remaining H cards containing this information then follow. If PC4 equals 1 the four quantities are to be read (in H groups of four) from the unformatted binary file with internal unit number indicated by PC5. If PC4 equals 2 the four quantities

are to be provided by the subroutine NSTRAT which is described in Section K3.2.4.

K3.1.3 The Variate Values. For each case in each stratum the user must provide the V variate values. These must be in the order of their relevant strata and exactly the number of cases as indicated by N must be present for each stratum. Furthermore the user should ensure that the values are weighted or unweighted according to the specified value of IV. (See previous section for a definition of N and IV).

If PC6 equals zero the user must provide a number of cards, specified by PC7 (maximum : 2) containing the format information concerning the cards from which the V values are to be read for each case in each stratum. The remaining cards containing all this information, then follow.

If PC6 equals one, each set of V is to be read from an unformatted binary file with internal unit number indicated by PC7. If PC6 equals two, each set of V is to be provided by the subroutine WINPUT which is described in Section 3.2.5.

K3.1.4 Computational Efficiency. The program may be instructed to disregard the input of data for certain strata for certain variates (see PC15 and Section K3.2.6); it may be instructed to use only population sums, or to disregard certain variates in the computation of certain functions (see PC16 and Section K3.2.7). On some systems these steps will be necessary, but on all systems they will enhance efficiency.

As a general rule computation will always be faster when population sums only are used, but also the writing of the

function subroutine will be easier and the numerical computation of derivatives will be more accurate. Thus, wherever possible, population sums should be used.

It was noted in the 'Worked Example' that different sets of variates may be used to calculate the same function. Where this is the case of the smallest set of variates will give the fastest solutions.

K3.2 The Main Program and the Subroutine

The user must always supply a main program and a double-precision function, F, in order to run the program. Several other subroutines will be necessary depending on the options specified on the Program Card. On some installations it will be necessary to provide dummy versions of these subroutines even if they are never called. The part of the program which has been supplied consists of three subroutines - PREPAR, GENVAR, and SWITCH - which must always be included in the source deck and which are referred to collectively as the 'core subroutines'.

K3.2.1 The Main Program. The main program performs five functions.

1 It begins the operation of the program. On some systems this is achieved by making the first line a 'PROGRAM' line; on others the system detects that it is a main program simply by the absence of a SUBROUTINE or FUNCTION statement at the beginning.

2 It reserves sufficient space for the dimensioned arrays which are to be used by all the subsequent subroutines.

The amount of space needed by the core subroutines is indicated below. If the space needed by these subroutines is not provided for in the main program, execution will terminate and one of the core subroutines will print out the necessary dimension limits. If any of the user-written subroutines require more space than is already reserved, the user must make allowances for this in the main program.

3 It can, beyond what is called for in the Problem Card, provide in advance data for any of the user-written subroutines. These could be read into interlocking COMMON areas not named COMMF or PREPGN for use at any subsequent stage.

4 It can be used to 'open' and 'close' files, rewind tapes, etc.

5 It calls the first core subroutine PREPAR.

The form of the main program is:

```
DIMENSION A(a),N(n)
DOUBLE PRECISION D(d)
      .
      .
      .
CALL PREPAR (A,a,N,n,D,d)
      .
      .
      .
STOP
END
```

As previously noted, a first 'PROGRAM' line may be necessary on some systems. 'A' is a real array, 'N' is an integer array, and 'D' is a double precision array; they are of dimension

a, n, and d respectively. These dimensions are actual numerical values which may be found by making a preliminary run with PC17 = 1 (for this preliminary run they may all be set equal to, say, 100). Alternatively, low bounds may be calculated by hand as follows:

If no temporary storage is called for (that is, all of PC10, PC11, PC12, PC13, PC14 are zero or blank) and if

H = number of strata

V = number of variates

R = total number of sample cases for all strata

K = number of sample cases in the longest stratum

then

$a \leq VR + K + H + 4V$

$n \leq 2H + V$

$d \leq VH + 2V$

These bounds need only be exceeded if the user has written subroutines which will use more dimensioned-array space. If population sums are used, or if the optional printouts are not called for, or if tempory storage on internal units is called for, then much smaller dimension bounds may be used. Exact bounds in these cases may be found by consulting Section K4 - 'The Problem of Insufficient Storage Space'.

K3.2.2 The Function F. The user must always provide a double precision function F defining the L functions under investigation. Double precision is needed here, to ensure accuracy in the numerical computation of the derivatives, and should be used to the fullest extent in defining F.

The form of the function F is:

```
      DOUBLE PRECISION FUNCTION F (T,NT,S,MS,NS,R,NR)
      DOUBLE PRECISION T(NT),S(MS,NS),R(NR)
      COMMON/COMMF/L
            .
            .
            .
      RETURN
      END
```

The arrays T, S and the common block COMMF should not be altered in a problem needing no temporary storage. If temporary storage is used consult Section K4. The common block contains L which is the number of the function to be evaluated. The function is to be defined in terms of population sums which are supplied in array T, or stratum sums which are supplied in array S, whichever is appropriate. Note that T is of dimension V and S is of dimension (H,V) that is, element S(I,J) will be the stratum sum for stratum I and variate J. Two possible pitfalls should be noted. First, if PC15 = 1, then for some pairs the stratum sum S(I,J) does not exist; if it appears in function F, its value must be set to zero. Secondly, for each function, and for each variable, the FUNCTION F must use stratum sums, population sums, or neither, according to what is called for by PC16 and SUBROUTINE NSUBFV.

K3.2.3 The Subroutine VARNCE. The user need supply this subroutine only if, for some strata, the internally-provided simple random sampling formula is not appropriate. The subroutine is called for each stratum in order to provide the variance of the U-statistics.

The form of the subroutine VARNCE is

```
SUBROUTINE VARNCE (T,M,I,N,FR,Q,INV)
DIMENSION T(M)
      .
      .
      .
RETURN
END
```

The array T consists of the U-statistics for the particular stratum under consideration. If the user indicated that the input data were unweighted, (see Section K3.1.2) the U-statistics will now be weighted. The other variables are:

M : the dimension of array T

I : the stratum under consideration

N : the stratum size

FR : the stratum sampling fraction

INV : an internal unit number needed only if temporary storage is used.

The variance contribution for each stratum is evaluated in terms of the U-statistics and placed in variable Q.

If temporary storage has been called for in the Problem Card, consult Section K4.

K3.2.4 Subroutine NSTRAT. The user need supply this subroutine only if PC4 equals 2. The subroutine must supply the four quantities N, FR, NT and IV for each stratum (see Section K3.1.2).

The form of the subroutine NSTRAT is

```
SUBROUTINE NSTRAT (I,N,FR,NT,IV)
      .
      .
      .
RETURN
END
```

The variable I holds the number of the stratum for which the four quantities are needed. One way of achieving this would be to read in or calculate the M sets of four quantities in the main program and transfer them to NSTRAT in a common block.

K3.2.5 Subroutine WINPUT. The user need supply this subroutine only if IS = 2. The subroutine must supply, in turn, a set of variate values for each case in each stratum (see Section K3.1.3).

The form of the subroutine WINPUT is;

```
SUBROUTINE WINPUT (W,M)
DIMENSION W(M)
      .
      .
      .
RETURN
END
```

The dimension of array W, if no temporary storage is called for, will be V, the number of variates. If temporary storage is called for, see Section K4. One way of arranging this subroutine would be to read in the variate values in the main program, transfer them to WINPUT in a common block, and read from this common block into array W the appropriate variate values of each case.

One extra integer variable, say K, included in the common block could be used to keep track of which was the appropriate case simply by incrementing K by one each time WINPUT is called.

K3.2.6 Subroutine NSUBHV. The user need supply this subroutine only if PC15 equals one. The subroutine must indicate to the calling subroutine, for each pair of stratum I and variate J, whether variate J is defined in stratum I.

The form of subroutine NSUBHV is:

```
SUBROUTINE NSUBHV (I,J,IHV)
      .
      .
      .
RETURN
END
```

The subroutine carries out its purpose by returning for each pair of stratum I and variate J the value IHV = 1 if variate J does not appear for stratum I and IHV = 0 otherwise. Note that for any pair for which IHV = 1, dummy variates must still be supplied for each of the cases in stratum I.

The data supplied in this subroutine should not be found by reading in fresh input from unit 5. One way of implementing the subroutine would be to read in the relevant data in the main program and transfer it to NSUBHV through a common block

K3.2.7 Subroutine NSUBFV. The user need supply this subroutine only if PC16 equals 1. The subroutine must indicate to the calling subroutine, for each pair of variate J and function L, whether;

1 the stratum sum of variate J is involved in the computation of function L;

2 the variate J is not involved at all in the computation of function L;

3 the population sum of variate J is involved in the computation of function L.

The form of the subroutine NSUBFV is

```
SUBROUTINE NSUBFV (L,J,IFV)
      .
      .
      .
RETURN
END
```

The value of IFV is to be set to

(a) zero if case (1) above holds

(b) one if case (2) above holds

(c) two if case (3) above holds.

The values of IHV should not be determined by reading-in fresh input from internal unit 5. One way of implementing this subroutine would be to read the required values of IHV into the main program and transfer them to NSUBFV in a common block.

K3.3 The Data

As indicated in the previous two sections, there are various types of data to be communicated to the program and there are several options about how each is communicated. There are five types of data input:

1 The control and definition information contained on the Problem Card.

2 The information describing each stratum.

3 The variate values.

4 Information concerning irrelevant strata and descriptors.

5 Sundry constants, weights, etc. that might be needed.

The information on the Problem Card is always expected in card form (or equivalently, on unit five). The stratum information may also be provided in card form, in which case it must be preceded by a format card, or on an unformatted binary file, or through a subroutine NSTRAT which may itself read cards or could alternatively read the information from a unit other than five (see PC4). The variate values may be read in using the same options as those for the stratum descriptors (see PC6). Information concerning irrelevant strata and variates is *not* to be read in from cards by subroutines NSUBRV and NSUBHV. This information may be read in from cards by the main program and transferred to the subroutines by common blocks, or it could be read from a unit other than five. Sundry weights, constants, etc. would be best read in from cards by the main program and then transferred to the relevant subroutines by common blocks, although it would be possible to read than from units other than five. If the user does decide to use auxiliary units care must be taken not to use any units needed for temporary storage (see PC10 to PC14) and to rewind them where relevant.

Data in card form (or equivalently on unit five) are called for in the following order:

1 Any data called for by the main program (see above)

2 The Problem Card (see K3.1)

3 If PC4 = 0, a card giving the format for the following stratum descriptors.

4 If PC4 = 0, or if NSTRAT reads cards, H cards supplying the stratum descriptors.

5 If PC6 = 0, one or two cards (whichever is indicated by PC7) giving the format for the following variate values.

6 If PC6 = 0, or if WINPUT reads cards, the user must here supply the cards which give the V variate values. There will be one or two cards per case (as indicated by PC7) and the cases are to be in order by strata.

K3.4 Printed Output

In order, the printed output is:

1 The required minimum dimensions for arrays A, N and D of the main program, based on the entries of the Program Card and the stratum descriptors (see K3.2.1). If PC17 = 1, the program stops after printing this information.

Then for each function:

2 If PC9 = 1, a line giving information for each computed derivative. These are grouped by stratum if stratum sums are being considered; if both stratum and population sums are being considered, the results for population are given for stratum one. The printed

information on each line is

(a) the number of the variate involved in the differentiation

(b) the number of iterations involved in the numerical procedure (** denotes the maximum of 8)

(c) the value of the derivative

(d) the final increment used in the numerical procedure.

If a sum for differentiation is less than 10^{-20} in absolute value, the derivative is set to zero, and this fact is indicated in the printed line.

3 If PC8 = 1, the 'U-statistics' are printed, grouped by function and stratum. These are printed in E11.6 format, 10 to a line.

4 The results, consisting of:

(a) the number of the function

(b) the estimated value of the function

(c) the estimated variance of the estimator of the function

(d) the estimated standard deviation (i.e. the square root of (c))

(e) the coefficient of variation (i.e. the standard deviation divided by the estimated value of the function.

K4 The Problem of Insufficient Storage Space

In certain circumstances the amount of storage space indicated in Section K3.2 will be larger than that available on particular systems. To cope with this a series of internal units may be used to reduce the need for large dimensioned arrays. Ignoring possible extra space needed by the user-written subroutines, the upper bounds to data storage space are:

1	floating-point	$a \leq VR + K + H + 4V$
2	integer	$n \leq 2H + V$
3	double precision	$d \leq VH + 2V$ (i.e. 2d words)

where

H = number of strata

V = number of variates

R = total number of sample cases for all strata

K = number of sample cases in the largest stratum.

These upper bounds may be much higher than those actually needed in any particular situation. The user may make a preliminary run (set PC17 to 1), in which case the program will calculate the appropriate upper bounds and then stop. If these upper bounds are within the available storage, the user should skip the remainder of this section

If the calculated upper bounds are too high, the user may try each of the following strategies in turn.

K4.1 Strategy One - Discard Variate Values

Generally the most troublesome storage problem is represented by the term 'VR' in the floating-point storage space. This is reduced to 1 if the user sets PC10 to 1 (format 11) and provides for a temporary storage file on unit 20. If more

temporary storage space is then required, try Strategy Two.

K4.2 Strategy Two-Reduce Number of Cases

In order to reduce the number of cases to be transferred to subroutine VARNCE, the user may place a value K* into PC11 (format I5). This will ensure that the U-statistics are transferred to subroutine VARNCE in sets of K* rather than sets of K. A temporary storage file must be provided on internal unit 29, and the size of K* is left to the discretion of the user (it should, of course, be smaller than K). This strategy reduces the second term in the floating-point storage space to K*.

If this strategy is used, the subroutine VARNCE must be altered to allow the reading of U-statistics from unit 29 (which is referred to internally as INV). The first line of the subroutine is: (see Section K3.2.3).

```
SUBROUTINE VARNCE (T,M,I,N,FR,Q,INV)
```

In this statement, N is the total stratum size (which, for the largest stratum, corresponds to K), M is the size of the array T which holds the U-statistics (and which will equal K* if K < N, and N otherwise), and INV is the unit on which the U-statistics are stored. The subroutine must be written so as to expect to find the U-statistics in array T if M ≥ N, or to read the U-statistics from unit INV in A + 1 sets of M, if M < N, where

$$A = \left[\frac{N}{M}\right]$$

= the largest integer smaller than $\left(\frac{N}{M}\right)$

In the latter case, the last set of U-statistics will contain B = N - MA values rather than M values. This is a rather complicated procedure but a series of statements such as the following will accomplish it.

```
      SUBROUTINE VARNCE (T,M,I,N,FR,Q,INV)
      DIMENSION T(M)
      L = 0
            .
            .                          Initialize variables etc.
            .
      ND = (N - 1)/M + 1
      DO 2 J = 1,ND
      IF (ND.EQ.1) GO TO 1
      READ (INV) T
1     DO 2 K = 1,M
      L = L + 1
      IF (L.GT.N) GO TO 3
            .
            .                          Calculate the variance
            .                          contribution from each
                                       set of M U-statistics
2     CONTINUE
3     Q = .....                        Put total variance into Q
      RETURN
      END
```

K4.3 Strategy Three - Discard Stratum Variate Sums

In order to reduce the number of variate sums transferred to function F in dimensioned arrays, the user may set PC12 to 1, and provide a temporary storage file on internal unit 21. This will reduce the first term in the double-precision storage space, VH to 1. Note that if only population sums are to be considered for *all* variates (i.e. if either PC16 = 2, or PC16 = 1 and all values of IHV are either 1 or 2) then this strategy is irrelevant.

If this strategy is employed the function F must be altered to allow for the reading of variate totals from an internal unit. The first two lines of the function must be

```
DOUBLE PRECISION FUNCTION F (T,NT,S,MS,NS,R,N.)
COMMON/COMMF/L,INP,I
```

The variable 'L' is, as before, the number of the function to be evaluated. The variable 'I' is the number of the stratum which has had its variate totals adjusted for the purpose of numerical differentiation; the values of these variate totals are contained in array R (of dimension V).

The variable 'INV' contains the number of the internal unit on which *all* the non-adjusted variate totals reside; these are in H groups of V. In order to understand the procedure described below the user must realize that the function F is used for two different purposes: first, it is called several times in the process of numerical differentiation; second, it is called to provide an estimate of the function which is printed as part of the output. The information regarding which purpose is appropriate is conveyed to the function using variable 'I':

1 if I > 0, the function is needed as part of the numerical differentiation process

2 if I = 0, the function is needed for the final estimate.

Case 1: I > 0. In this case the values of the V variate sums for stratum I are contained in array R. In the first step they should be manipulated in the appropriate way and the results stored elsewhere.

The pointer on the internal unit INV will now be located at the first variate sum for the first case in stratum I + 1. The set of variate sums for stratum I + 1 are then read in from unit INV, the appropriate manipulation carried out, and the results stored elsewhere. This process is repeated for stratum I + 2, I + 3 and so on up to stratum H. Unit INV is then rewound and the same process repeated for strata 1, 2 ... I - 1. Note that the set of variate sums for stratum I is not to be read from unit INP.

Case 2: I = 0. In this case the H sets of variate sums are to be read from unit INP and handled appropriately. The pointer on unit INP will be pointing at the first case of stratum I.

The following set of statements would be one way of implementing the above procedure. For simplicity it is assumed that there is only one function to be evaluated by function F; h represents the number of strata.

```
      DOUBLE PRECISION FUNCTION F (T,NT,S,MS,NS,R,NR)
      DOUBLE PRECISION T(NT),S(MS,NS),R(NR)
      COMMON/COMMF/L,INP,I
      IF (I.EQ.0) GO TO 1
           .
           .
           .
           .
           .
      IF (I.EQ.h) GO TO 3
    1 IP = I + 1
      DO 2 K = IP,h
      READ (INP) R
           .
           .
           .
           .
           .
```

Variate totals for stratum I are contained in array R. Relevant manipulations are carried out and the results stored elsewhere

Variate totals for strata I + 1, I + 2 ... h are read in, one stratum at a time, manipulated, and the results stored.

```
2     CONTINUE
3     REWIND IND
      IF (I.LE.1) GO TO 5
      IP = I - 1
      DO 4 K = 1,IP
      READ (INP) R
            .
            .
            .
            .
            .
4     CONTINUE
5     F = ..........
      RETURN
      END
```

Variate totals for strata 1, 2 ... I - 1 are read in, one stratum at a time, manipulated, and the results stored.

Calculate F using the stored values of the variate totals for all strata

K4.4 Strategy Four - Delete Stratum Descriptors

In order to eliminate the need to store stratum descriptors in dimensioned arrays, the user may place a 1 in PC13 and provide temporary storage in unit 19. This will have the effect of replacing the term 'H' in the floating-point total by 1, and of replacing the term 2H in the integer total by 2.

K4.5 Strategy Five - The Last Resort

If after resorting to all the above measures there is still insufficient storage space, the user may enter in PC14 the value V* (format I5) which will cause each set of V variates to be dealt within sets of size V*. The user will need to provide temporary storage on units 22 through 27 and also 28 if both

(a) population sums are used, and

(b) not all variables are defined for all strata.

This will reduce the terms 4V, V and 2V in the floating-point, integer, and double precision totals to 4V*, V* and 2V* respectively.

The implementation of this strategy involves the alteration of subroutine WINPUT and of function F. First, some notation is necessary; let V and V* be as above, then let

$$A = \left[\frac{V}{V^*}\right]$$

= the largest integer smaller than V ÷ V*

$$B = V - AV^*$$

$$C = \begin{cases} A \text{ if } B = 0 \\ A + 1 \text{ if } B > 0 \end{cases}$$

The subroutine WINPUT must be written so as to supply to the calling program, for each case, C subsets of V* variates, with only B variates contained in the last subset if B > 0. The modifications to subroutine WINPUT are basically the same as those made for subroutine VARNCE in Strategy Two (see K3.2.5 and K4.2).

The alterations to subroutine F differ depending upon whether population sums or stratum sums are involved. In either case the user must provide a common statement,

```
COMMON/COMMF/L,INP,I,INT,M
```

If population sums are involved, the array of population sums is given in C sets of V*. Just as for Strategy Four, the variable I determines the use to which F is to be put. If I = 0 the population sums are read from unit INT in C sets of V* and all are used to calculate F. If I > 0, then the array T will contain the V* population sums for set M; these are manipulated and then, as in Strategy Four, the remaining sets M + 1, M + 2 ... C are read in and manipulated, the unit INT is rewound, and

the sets 1, 2 ... M - 1 are also read in and manipulated. The Fortran example in K4.3 may be modified as follows to implement this procedure. Let c be the number of subsets (i.e. c ≥ C).

```
      DOUBLE PRECISION FUNCTION F(T,NT,S,MS,NS,R,NR)
      DOUBLE PRECISION T(NT),S(MS,NS),R(NR)
      COMMON/COMMF/L,INP,I,INT,M
      IF (I.EQ.0) GO TO 1
            .                                  } Manipulate the variate
            .                                    sums for subset M
            .
      IF (M.EQ.c) GO TO 4
      IP = M + 1
      GO TO 2
1     IP = I + 1
2     DO 3 K = IP,c
      READ (INT) T
            .                                  } Manipulate the variate
            .                                    sums for subsets M + 1,
            .                                    M + 2, ... c
3     CONTINUE
4     REWIND INT
      IF (I.EQ.0.OR.M.EQ.1) GO TO 6
      IP = M - 1
      DO 5 K = 1,IP
      READ (INT) T
            .                                  } Manipulate the variate
            .                                    sums for subsets 1, 2
            .                                    ... M - 1
5     CONTINUE
6     F = .....                                  Calculate F
      RETURN
      END
```

If stratum sums are involved, a procedure combining both the alterations to F described in Strategy Three, and those above for the case of population sums must be implemented. The procedure is described below.

The array of stratum sums is given in H sets of V (one for each stratum), each of these broken into C subsets of V*. For I > 0 the current values of R are for set I, subset M. The user may use these if desired, and (whatever the value of I) must:

1 for M < C, read from INP, sets of V* values, one at a time, and do the corresponding calculations for subsets M + 1, M + 2 ... C for set I;

2 for I < H, read C subsets and do likewise for each subset, for each of the strata I + 1, I + 2 ... H;

3 Rewind INP

4 for I > 1, read C subsets and do likewise for each subset, for each of strata 1, 2 ... I - 1;

5 for I > 0 and M > 1 read subsets 1, 2 ... M - 1 and do likewise for stratum I.

K4.6 Strategy Six - The Ultimate Resort

If, after attempting all the above strategies, there is still insufficient storage space, the user is advised to try a trip to Tahiti. It may not solve the storage problems, but from a new perspective the user may well decide it does not matter.

K4.7 Summary

The situation is summarized in Table K4.1.

Table K4.1 Summary of Strategies for Use of Temporary Storage

Strategy	Indicator on Problem Card	Columns	User must provide unit	Space saved	Problems
1	PC10=1	21	20	VR-1	none
2	PC11=K*	22-26	29	K-K*	alterations to VARNCE
3	PC12=1	27	21	VH-1	alterations to F
4	PC13=1	28	19	3H-3	none
5	PC14=V*	29-33	22-28	6V-6V*	alterations to WINPUT alterations to F

K5 Sample Inputs and Outputs

These examples, provided by the program's authors, are deliberately contrived to illustrate features of the program.

Example A

Main Deck (see K3.2.1):

```
      DIMENSION A(60),N(60)
      DOUBLE PRECISION D(60)
      DIMENSION T(1),NB(1)
      COMMON/ASAMP/NB,T
      READ(5,60) (NB(I),I=1,1)
   60 FORMAT(10I8)
      READ(5,63) (T(I),I=1,1)
   63 FORMAT (20F4.0)
      DIMENSION WT(5,3)
      COMMON/AW/WT,K
      K = 0
      DO 51 L=1,3
      READ(5,600) WT(1,L),WT(2,L)
  600 FORMAT(2F1.0)
      WT(3,L) = WT(1,L)**2
      WT(4,L) = WT(2,L)**2
   51 WT(5,L) = WT(1,L)*WT(2,L)
      CALL PREPAR(A,60,N,60,D,60)
      STOP
      END
```

Subroutines:

(a) (see K3.2.5)

```
      SUBROUTINE WINPUT(W,N)
      DIMENSION WT(5,3),W(N)
      COMMON/AW/WT,K
      K = K + 1
      DO 54 J=1,5
54    W(J) = WT(J,K)
      RETURN
      END
```

(b) (see K3.2.4)

```
      SUBROUTINE NSTRAT(I,N,FR,NT,IV)
      DIMENSION NB(1),T(1)
      COMMON/ASAMP/NB,T
      N = NB(I)
      FR = T(1)
      NT = 0
      IV = 1
      RETURN
      END
```

(c) (see K3.2.2)

```
      DOUBLE PRECISION FUNCTION F(XT,NT,XS,MS,NS,XR,NR)
      DOUBLE PRECISION XT(NT),XS(MS,NS),XR(NR)
      F = (XT(5) - XT(1)*XT(2)/75.)/
     1DSQRT((XT(3) - XT(1)**2/75.)*(XT(4) - XT(2)**2/75.))
      RETURN
      END
```

(Value of 'L' not needed as there is only 1 function).

Data Deck:

Card 1: Column 8 = 1 (see K3.2.1 and K3.2.4)

Card 2: Columns 1-4 = 0.04 (see K3.2 and K3.2.4)

Card 3: Column 1 = 1 (see K3.2 and K3.2.5)

Column 2 = 2

Card 4: Column 1 = 0 (see K3.2 and K3.2.5)

Column 2 = 3

Card 5: Column 1 = 2 (see K3.2 and K3.2.5)

Column 2 = 8

Card 6: Column 4 = 1 (PC1)

Column 8 = 1 (PC2)

Column 12 = 5 (PC3)

Column 13 = 2 (PC4)

Column 16 = 2 (PC6)

Column 35 = 2 (PC16)

Printed Output:

```
DIMENSION LIMITS ARE 33  7  7
FUNCTION  1  .69537524+00  .64704124-01  .25437009+00  .36685776+00
```

Example B (Same problem as Example A)

Main Deck (see K3.2.1):

```
      DIMENSION A(60),N(60)
      DOUBLE PRECISION D(60)
      CALL PREPAR(A,60,N,60,D,60)
      STOP
      END
```

Subroutines:

(a) (see K4.2 and K3.2.3; note that this is simple random sample formula)

```
      SUBROUTINE VARNCE(T,M,I,N,FR,Q,INU)
      DIMENSION T(M)
      L = 0
      Q = 0.
      B = 0.
      ND = (N - 1)/M + 1
      A = N
      DO 1 J=1,ND
      IF (ND.EQ.1) GO TO 3
      READ(INU) T
    3 DO 1 K=1,M
      L = L + 1
```

```
      IF (L.GT.N) GO TO 2
      S = L
      H = B + T(K)
      IF (L.EQ.1) GO TO 1
      Q = Q + (S*T(K) - B)**2/(S*(S - 1.))
1     CONTINUE
2     Q = (1. - FR)*A*Q/(A - 1.)
      RETURN
      END
```

(b) (see K3.2.2 and K4.4)

```
      DOUBLE PRECISION FUNCTION F(XT,NT,XS,MS,NS,XR,NR)
      DOUBLE PRECISION XT(NT),XS(MS,NS),XR(NR)
      COMMON/COMMF/L,INP,I,INT,M
      DOUBLE PRECISION XP(6)
      J = 2*M
      IF (M.EQ.0) GO TO 1
      XP(J-1) = XT(1)
      XP(J) = XT(2)
      IF (M.EQ.3) GO TO 4
1     MP = M + 1
      DO 3 K = MP,3
      READ(INT) XT
      XP(J+1) = XT(1)
      XP(J+2) = XT(2)
3     J = J + 2
4     REWIND INT
      J = 0
      IF (M.LE.1) GO TO 6
      MP = M - 1
      DO 5 K=1,MP
      READ(INT) XT
      XP(J+1) = XT(1)
      XP(J+2) = XT(2)
5     J = J + 2
6     F = (XP(5) - XP(1)*XP(2)/75.)/
     1DSQRT((XP(3) - XP(1)**2/75.)*(XP(4) - XP(2)**2/75.))
      RETURN
      END
```

Data Deck:

Card 1: Column 4 = 1 (PC1)

Column 8 = 1 (PC2)

Column 12 = 5 (PC3)

Column 19 = 1 (PC8)

Column 20 = 1 (PC9)

Column 21 = 1 (PC10)

Column 26 = 2 (PC11)

Column 28 = 1 (PC13)

Column 33 = 2 (PC14)

Column 35 = 2 (PC16)

Card 2: Columns 1-15 = (I8,F4.0,I2,I1) (see K3.1.2)

Card 3: Column 8 = 3

Columns 13-14 = 75

Column 15 = 3

Card 4: Columns 1-7 = (2F2.0) (see K3.1.3)

Cards 5-14: 0101

0101

0100

0003

0009

0000

0208

0464

1600

Printed Output:

```
DIMENSION LIMITS ARE 11  4  4

DERIVATIVES FOR FUNCTION  1

     1   3   -.85205026-02    .27159816-05

     2   3   -.12800773-02    .15624528-04

     3   3   -.69337523-02    .41706269-05

     4   3   -.53336558-03    .61858479-04

     5   3    .55470018-02    .14210732-04

U-VALUES FOR FUNCTION  1,STRATUM  1

-.28017+00    -.216915+00

 .232831-07

FUNCTION 1  .69357524+00  .64704124-01  .25437009+00  .36685776+00
```

Note: Internal units 19, 20, 22, 23, 24, 25, 26, 27 and 29 are all used.

Example C

Main Deck: same as for Example B

Subroutines:

(a) (see K3.2.6.)

```
SUBROUTINE NSUBHV (I,J,IHV)
IHV = MIN0(I-1,J-1)
RETURN
END
```

(b) (see K3.2.7)

```
SUBROUTINE NSUBFV (I,J,IFV)
IFV = MIN0 (I-1, J-1)
RETURN
END
```

(c) (see K3.2.2)

```
DOUBLE PRECISION FUNCTION F(XT,NT,XS,MS,NS,NR)
DOUBLE PRECISION XT(NT),XS(MS,NS),XR(NR)
COMMON COMNF/I
F = XS(2,1)/XS(1,1)
```

```
      IF (L.EQ.2) RETURN
      F = F*XS(1,2)
      RETURN
      END
```

Data Deck:

Card 1: Column 4 = 2

Column 8 = 2

Column 12 = 2

Column 34 = 1

Column 35 = 1

Card 2: Columns 1-13 = (I1,F1.0,2I1)

Card 3: Column 1 = 2

Column 3 = 4

Card 4: same as Card 3

Card 5: Columns 1-7 = (2F1.0)

Cards 7-10: 24

26

00

20

Printed Output:

```
DIMENSION LIMITS ARE 19  9  6
FUNCTION  1  .50000000+01  .13000000+02  .36055513+01  .72111025+00
FUNCTION  2  .50000000+00  .12500000+00  .35355339+00  .70710678+00
```

K6 Listing of the Program

```
      SUBROUTINE PREPAR(A,IA,B,IB,D,ID)                                 00000010
      DIMENSION A(IA),N(IN),FMT(12)                                     00000020
      DOUBLE PRECISION D(ID)                                            00000030
      COMMON/PREPON/MA,MF,MV,JMU,KMU,JU,JF,KM,KS,MMV,MFV,MEVT,IMS       00000040
      READ(5,60) MM,MF,MV,JMS,IMS,JMU,KMU,JU,JF,KU,KM,KM,KS,MMV,MMV,MFV,00000050
     1MFR                                                               00000060
   60 FORMAT(3I4,2(I1,I2),3I1,I5,2I1,I5,3I1)                            00000070
      MM = 0                                                            00000080
      IB3 = 1                                                           00000090
      IMS=19                                                            00000100
      IF (JMS .GT. 0) GO TO 6                                           00000110
      READ(5,61) FMT                                                    00000120
   61 FORMAT(12A6)                                                      00000130
    6 DO 1 I=1,MM                                                       00000140
      IF (JMS - 1) 2,3,4                                                00000150
    4 CALL PSTART(I,MM,FR,MT,IV)                                        00000160
      GO TO 5                                                           00000170
    3 READ(IMS) MM,FR,MT,IV                                             00000180
      GO TO 5                                                           00000190
    2 READ(5,FMT) MM,FR,MT,IV                                           00000200
    5 IF (JU .EQ. 0 .AND. IV .LT. 2) GO TO 9                            00000210
      IB3 = MAX0(IB3,MM)                                                00000220
    9 IF (MT .EQ. 0) GO TO 7                                            00000230
      MM = MT                                                           00000240
      B = MM                                                            00000250
      FR = B/MM                                                         00000260
    7 IF (KS .EQ. 0 .OR. MFR .GT. 0) GO TO 8                            00000270
      WRITE(IMS) MM,FR,IV                                               00000280
    8 MR = MR + MM                                                      00000290
      IF (KS .GT. 0 .OR. MFR .GT. 0) GO TO 1                            00000300
      A(I) = FR                                                         00000310
      N(I) = MM                                                         00000320
      N(I + MM) = IV                                                    00000330
    1 CONTINUE                                                          00000340
      IF (KS .EQ. 0 .OR. MFR .GT. 0) GO TO 24                           00000350
      REWIND IMS                                                        00000360
   24 IB1 = 1                                                           00000370
      IF (KS .GT. 0) GO TO 11                                           00000380
      IB1 = MM                                                          00000390
   11 I2 = 1 + IB1                                                      00000400
      J2 = 1 + IB1                                                      00000410
      J3 = J2 + IB1                                                     00000420
      IB2A = 1                                                          00000430
      IB2B = 1                                                          00000440
      IF (MMU .GT. 0) GO TO 12                                          00000450
      IB2A = MR                                                         00000460
      IB2B = MV                                                         00000470
   12 I3 = I2 + IB2A+IB2B                                               00000480
      IF (KM .EQ. 0 .OR. KM .GT. IB3) GO TO 13                          00000490
      IB3 = KM                                                          00000500
```

```
13 I4 = I3 + ID3
14 KD1A = 1
   KD1B = 1
   NEVT = NFV
   IF (NFV .NE. 1) GO TO 15
   NA = 1
   NB = 1
   DO 16 L=1,NF
   DO 16 J=1,NV
   CALL NSUBFV(L,J,IFV)
   NA = MIN0(NA,IFV)
   NB = MIN0(NB,IFV)
16 CONTINUE
   IF (NA .EQ. 0 .AND. NB .EQ. 2) GO TO 15
   IF (NA .EQ. 1) GO TO 17
   NEVT = 0
   GO TO 15
17 NEVT = 2
15 IF (NFVT .EQ. 2 .OR. KH .EQ. 1) GO TO 18
   KD1A = NH
   KD1B = NV
18 K2 = 1 + KD1A*KD1B
   NVH = NV
   IF (NVV .EQ. 0) GO TO 19
   NVH = NVV
19 I5 = I4 + NVH
   I6 = I5 + NVH
   I7 = I6 + NVH
   JT = J3 + NVH - 1
   KD2 = 1
   IF (NFVT .EQ. 0) GO TO 20
   KD2 = NVH
20 K3 = K2 + KD2
   KD3 = 1
   IF (NFVT .EQ. 2 .OR. KH .EQ. 0) GO TO 21
   KD3 = NVH
21 KT = K3 + KD3 - 1
   ID7 = 1
   IF (NFVT .EQ. 0) GO TO 22
   IF (NFVT .EQ. 2 .AND. NAV .EQ. 0) GO TO 22
   ID7 = NVH
22 IT = I7 + ID7 - 1
   WRITE(6,60) IT,JT,KT
60 FORMAT(1X,21HDIMENSION LIMITS ARE 3(1X,I5))
   IF (NER .GT. 0 .OR. IT .GT. IA) GO TO 23
   IF (JT .GT. IW .OR. KT .GT. ID) GO TO 23
   CALL GENIMR(A,ID1,A(I2),ID2A,ID2B,A(I3),ID3,A(I4),A(I5),A(I6),NVH,
  1A(I7),ID7,W,W(J2),W(J3),D,KD1A,KD1B,D(K2),KD2,D(K3),KD3)
23 RETURN
   END
```

```
      SUBROUTINE GENVAR(TT,ID1,WT,ID2A,ID2B,U,KU,W,SU,YY,NVH,YZ,ID7,
     1NBT,IVT,ID,XP,KD1A,KD1B,XT,KD2,XX,KD3)
      DIMENSION FMTW(24),TT(ID1),WT(ID2A,ID2B),U(KU),W(NVH),SW(NVH),
     1YY(NVH),YZ(ID7),NBT(ID1),IVT(ID1),ID(NVH)
      DOUBLE PRECISION XP(KD1A,KD1B),XT(KD2),XX(KD3),F,AP,AL
      DOUBLE PRECISION R1,R2,R3,P1,PIT
      COMMON/PPERGN/NN,NP,NV,JNW,KNW,JU,JP,KM,KS,NHV,NFV,NFVI,INS
      COMMON/COMNP/E,INP,I,INTA,JD
      DATA INW,INSA,INS5,INTB,INO,INT,INZ,INU/
     120,22,23,25,26,27,28,29/
      NVD = (NV - 1)/NVH + 1
      INP=21
      INTA=24
      NVR = NV - NVH*(NVD - 1)
      DATA AD,EXZ,EX1,AVAL/1.E5,1.E6,1.E8,1.E-20/
      IF (JNW .GT. 0) GO TO 405
      NTH = 12*MAX0(KNW,1)
      READ(5,63) (FMTW(I),I=1,NTH)
      WRITE(6,63) (FMTW(I),I=1,NTH)
   63 FORMAT(12A6)
  405 KY = 0
      DO 110 I=1,NN
      IF (KS .EQ. 0) GO TO 404
      READ(INS) NSAMP,FRAC,IVAR
      GO TO 4031
  404 NSAMP = NBT(I)
      FRAC = TT(I)
      IVAR = IVT(I)
 4031 JC = 0
      DO 1151 JD=1,NVD
      NVC = MIN0(NVH,NV-JC)
      DO 1153 J=1,NVC
      JC = JC + 1
      IF (I .GT. 1 .OR. NFVT .EQ. 0) GO TO 1155
      XT(J) = 0.
 1155 ID(J) =JC
      IF (NHV .EQ. 0) GO TO 1153
      CALL NSUBNV(I,JC,INV)
      IF (INV .EQ. 0) GO TO 1153
      ID(J) = 0
 1153 CONTINUE
      IF (NVD .EQ. 1) GO TO 1151
      WRITE(INO) ID
      IF (I .GT. 1 .OR. NFVT .EQ. 0) GO TO 1151
      WRITE(INTA) XT
 1151 CONTINUE
      IF (NVD .EQ. 1) GO TO 1154
      REWIND INO
      IF (I .GT. 1 .OR. NFVT .EQ. 0) GO TO 1154
      REWIND INTA
 1154 DO 12 K=1,NSAMP
      KY = KY + 1
```

```
      JC = 0                                                            00001530
      DO 121 JD=1,NVD                                                   00001540
      IF (JHU - 1) 4051,4052,4053                                       00001550
 4053 CALL UINPUT(U,NVH)                                                00001560
      GO TO 406                                                         00001570
 4052 READ(INU) U                                                       00001580
      GO TO 406                                                         00001590
 4051 READ(5,FMTU) U                                                    00001600
  406 NVC = MINO(NVH,NV-JC)                                             00001610
      IF (IDSK .GT. 1) GO TO 4060                                       00001620
      WRITE(INU) U                                                      00001630
      GO TO 4061                                                        00001640
 4060 DO 4062 J=1,NVC                                                   00001650
      JC = JC + 1                                                       00001660
 4062 UT(KY,JC) = U(J)                                                  00001670
 4061 IF (NVD .EQ. 1) GO TO 1223                                        00001680
      READ(IND) ID                                                      00001690
      IF (K .EQ. 1) GO TO 1224                                          00001700
      READ(INSA) SU                                                     00001710
 1223 IF (K .GT. 1) GO TO 1221                                          00001720
 1224 DO 123 J=1,NVC                                                    00001730
  123 SU(J) = 0.                                                        00001740
 1221 DO 11 J=1,NVC                                                     00001750
      IF (ID(J) .EQ. 0) GO TO 11                                        00001760
      SU(J) = SU(J) + W(J)                                              00001770
   11 CONTINUE                                                          00001780
      IF (NVD .EQ. 1) GO TO 121                                         00001790
      WRITE(INSB) SU                                                    00001800
  121 CONTINUE                                                          00001810
      IF (NVD.EQ.1) GO TO 12                                            00001820
      REWIND IND                                                        00001830
      CALL SWITCH(INSA,INSB)                                            00001840
   12 CONTINUE                                                          00001850
      DO 125 JD=1,NVD                                                   00001860
      IF (NVD .EQ. 1) GO TO 126                                         00001870
      READ(IND) ID                                                      00001880
      READ(INSA) SU                                                     00001890
      IF (NFVT .EQ. 0) GO TO 126                                        00001900
      READ(INTP) XT                                                     00001910
  126 DO 114 J=1,NVH                                                    00001920
      IF (J .GT. NVR .AND. JD .EQ. NVD) GO TO 1270                      00001930
      IF (ID(J) .EQ. 0) GO TO 114                                       00001940
      IF (IVAR .EQ. 0 .OR. IVAR .EQ. 2) GO TO 1142                      00001950
      SU(J) = SU(J)/FRAC                                                00001960
 1142 IF (NFVT .EQ. 2) GO TO 11422                                      00001970
      IF (KK .EQ. 0) GO TO 11420                                        00001980
      XX(J) = SU(J)                                                     00001990
      GO TO 11421                                                       00002000
11420 XP(I,J) = SU(J)                                                   00002010
11421 IF (NFVT .EQ. 0) GO TO 114                                        00002020
11422 XT(J) = XT(J) + SU(J)                                             00002030
```

```
  114 CONTINUE                                                          00002070
 1270 IF (NVD .EQ. 1) GO TO 125                                         00002080
      IF (NPVT .EQ. 0) GO TO 127                                        00002090
      WRITE(INT6) XT                                                    00002070
  127 IF (NPVT .EC. 2 .OR. KH .EQ. 0) GO TO 125                         00002080
      WRITE(INP) XX                                                     00002090
  125 CONTINUE                                                          00002100
      IF (NPVT .EQ. 2 .OR. KH .EQ. 0) GO TO 1250                        00002110
      IF (NVD .GT. 1) GO TO 1250                                        00002120
      WRITE(INP) XX                                                     00002130
 1250 IF (NVD .EQ. 1) GO TO 110                                         00002140
      REWIND IND                                                        00002150
      REWIND INSA                                                       00002160
      IF (NPVT .EQ. 0) GO TO 110                                        00002170
      CALL SWITCH(INTA,INTB)                                            00002180
  110 CONTINUE                                                          00002190
      IF (INSA .GT. 1) GO TO 134                                        00002200
      REWIND INW                                                        00002210
      IF (KS .EQ. 0) GO TO 134                                          00002220
      REWIND INS                                                        00002230
  134 IF (NVD .EQ. 1 .OR. NPVT .EQ. 0) GO TO 131                        00002240
      REWIND INTA                                                       00002250
  131 IF (KH .EQ. 0) GO TO 132                                          00002260
      REWIND INP                                                        00002270
  132 DO 102 L=1,NF                                                     00002280
      I = 0                                                             00002290
      JD = 0                                                            00002300
      VFR = F(XT,KD2,XF,KD1A,KD1B,XX,KD3)                               00002310
      DMR=ABS(VFR)                                                      00002320
      A = 0.                                                            00002330
      KT = 0                                                            00002340
      DO 1 I=1,KH                                                       00002350
      IF (JF .EQ. 0) GO TO 1001                                         00002360
      IF (NPVT .LT. 2) GO TO 1002                                       00002370
      WRITE(6,84) L                                                     00002380
   84 FORMAT(/1X,23HDERIVATIVES FOR FUNCTION  I3)                       00002390
      GO TO 100;                                                        00002400
 1002 WRITE(6,8041 L,I                                                  00002410
  804 FORMAT(/1X,23HDERIVATIVES FOR FUNCTION  I3,10H, STRATUM  I3)      00002420
 1001 IF (I .EQ. 1) GO TO 1003                                          00002430
      IF (NPVT .EQ. 2 .AND. NAV .EQ. 0) GO TO 103                       00002440
 1003 JD = 0                                                            00002450
      IF (KH .EQ. 0 .OR. NPVT .EQ. 2) GO TO 1004                        00002460
      IF (NVD .GT. 1) GO TO 1004                                        00002470
      READ(INP) XX                                                      00002480
 1004 DO 120 JD=1,NVD                                                   00002490
      IF (NVD .EQ. 1 .OR. NPVT .EQ. 0) GO TO 128                        00002500
      READ(INT6) XT                                                     00002510
      IF (I .EQ. 1) GO TO 128                                           00002520
      READ(INZ) YZ                                                      00002530
  128 IF (KH .EQ. 0 .OR. NPVT .EQ. 2) GO TO 107                         00002540
```

```
      IF (NVD .EQ. 1) GO TO 129                                         00002550
      READ(INP) XX                                                      00002560
  129 NVC = MIN0(NVR,NV-JC)                                             00002570
      DO 101 J=1,NVC                                                    00002580
      JC = JC + 1                                                       00002590
      IF (NFV - 1) 16,15,14                                             00002600
   15 CALL NSUBFV(L,JC,IFV)                                             00002610
      IF (IFV .EQ. 1) GO TO 310                                         00002620
      CALL NSUBNV(I,JC,INV)                                             00002630
      IF (INV .EQ. 1) GO TO 310                                         00002640
      IF (IFV .EQ. 0) GO TO 16                                          00002650
   14 ID(J) = 1                                                         00002660
      IF (I .GT.1) GO TO 32                                             00002670
      AB = DABS(XT(J))                                                  00002680
      IF (AB .LT. AVAL) GO TO 31                                        00002690
      GO TO 17                                                          00002700
   32 YT(J) = YZ(J)                                                     00002710
      GO TO 101                                                         00002720
   16 ID(J) = 2                                                         00002730
      IF (KH .GT. 0) GO TO 161                                          00002740
      AB = DABS(XP(I,J))                                                00002750
      GO TO 17                                                          00002760
  161 AB = DABS(XX(J))                                                  00002770
   17 IF (AB .LT. AVAL) GO TO 31                                        00002780
      RIF = 2.*EX2/AB                                                   00002790
      R2F = EX1/DAB                                                     00002800
    2 IT = 0                                                            00002810
      IK = 0                                                            00002820
      DI = AB/AD                                                        00002830
      DIT = 100.*DI                                                     00002840
    9 IT = IT + 1                                                       00002850
      IF (IT .GT. 3) GO TO 30                                           00002860
      R1 = DI*R1F                                                       00002870
      IF (R1 .LT. .1) GO TO 3                                           00002880
      IF (ID(J) .EQ. 1) GO TO 173                                       00002890
      IF (KH .GT. 0) GO TO 172                                          00002900
      XP(I,J) = XP(I,J) + DI                                            00002910
      AP = F(XT,KD2,XP,KD1A,KD1B,XX,KD3)                                00002920
      XP(I,J) = XP(I,J) - 2.*DI                                         00002930
      AL = F(XT,KD2,XP,KD1A,KD1B,XX,KD3)                                00002940
      XP(I,J) = XP(I,J) + DI                                            00002950
      GO TO 171                                                         00002960
  173 XT(J) = XT(J) + DI                                                00002970
      AP = F(XT,KD2,XP,KD1A,KD1B,XX,KD3)                                00002980
      XT(J) = XT(J) - 2.*DI                                             00002990
      IF (NVD .EQ. 1) GO TO 1731                                        00003000
      READ(INTA) XT                                                     00003010
      XT(J) = XT(J) - DI                                                00003020
 1731 AL = F(XT,KD2,XP,KD1A,KD1B,XX,KD3)                                00003030
      XT(J) = XT(J) + DI                                                00003040
      IF (NVD .EQ. 1) GO TO 171                                         00003050
```

```
      READ(INTAB) XT                                                    00003060
      GO TO 171                                                         00003070
  172 XX(J) = XX(J) + DI                                                00003080
      AP = F(XT,KD2,XP,KD1A,KD1B,XX,KD3)                                00003090
      READ(INP) XX                                                      00003100
      XX(J) = XX(J) - DI                                                00003110
      AL = F(XT,KD2,XP,KD1A,KD1B,XX,KD3)                                00003120
      READ(INP) XX                                                      00003130
  171 R2 = R2F+DABS(AP-AL)                                              00003140
      IF (R2 .LT. .1) GO TO 8                                           00003150
      IF (R1 .LT. 10. .OR. R2 .LT. 10.) GO TO 7                         00003160
      RJ = DMIN1(R1,R2)                                                 00003170
      RJ = DSQRT(RJ)                                                    00003180
      GO TO 5                                                           00003190
    5 RJ = R1                                                           00003200
      GO TO 6                                                           00003210
    6 RJ = R2                                                           00003220
      IK = IK + 1                                                       00003230
      IF (IK .LT. 3) GO TO 8                                            00003240
   30 IT = 10000                                                        00003250
      GO TO 2                                                           00003260
    8 DI = DI/RJ                                                        00003270
      DI = DMIN1(DI,DIT)                                                00003280
      GO TO 9                                                           00003290
   31 IF (JP .EQ. 0) GO TO 310                                          00003300
      WRITE(6,602) JC                                                   00003310
  602 FORMAT(5X,I3,26H ZERO SUM, ZERO DERIVATIVE  )                     00003320
  310 YY(J) = 0.                                                        00003330
      ID(J) = 0                                                         00003340
      GO TO 101                                                         00003350
    7 YY(J) = (AP-AL)*.5/DI                                             00003360
      IF (JP .EQ. 0) GO TO 101                                          00003370
      WRITE(6,600) JC,IT,YY(J),DI                                       00003380
  600 FORMAT(5X,I3,1X,I2,2(1X,E15.8))                                   00003390
  101 CONTINUE                                                          00003400
      IF (NWD .EQ. 1) GO TO 1202                                        00003410
      WRITE(INY) YY                                                     00003420
      WRITE(IND) ID                                                     00003430
 1202 IF (I .GT. 1 .OR. NPVT .EQ. 0) GO TO 120                          00003440
      IF (NPVT .EQ. 2 .AND. NHV .EQ. 0) GO TO 120                       00003450
      IF (NWD .EQ. 1) GO TO 1204                                        00003460
      WRITE(INZ) YY                                                     00003470
      GO TO 120                                                         00003480
 1204 DO 1205 J=1,NVH                                                   00003490
 1205 YZ(J) = YY(J)                                                     00003500
  120 CONTINUE                                                          00003510
  103 IF (NS .EQ. 0) GO TO 1031                                         00003520
      READ(INS) NSAMP,FRAC,IVAR                                         00003530
      GO TO 1032                                                        00003540
 1031 NSAMP = NBY(I)                                                    00003550
      FRAC = YY(I)                                                      00003560
```

```
      IVAR = IVT(I)                                                     00003520
 1032 IF (NVD .EQ. 1) GO TO 1030                                        00003530
      REWIND INTA                                                       00003540
      REWIND IND                                                        00003550
      REWIND INY                                                        00003560
      IF (NFVT .EQ. 0) GO TO 1030                                       00003570
      IF( NFV. .EQ. 2 .AND. NHV .EQ. 0) GO TO 1030                      00003580
      REWIND IHZ                                                        00003590
 1030 B = 0.                                                            00003600
      KHU = 0                                                           00003610
      KVAR = IVAR/2                                                     00003620
      JVAR = IVAR - 2*KVAR                                              00003630
      IF (KVAR .EQ. 0 .OR. NSAMP .LE. KU) GO TO 1034                    00003640
      KHU = 1                                                           00003650
 1031 Q = 0.                                                            00003660
      ICT = 0                                                           00003670
      IF (JU .EQ. 0) GO TO 201                                          00003680
      WRITE(6,803) L,I                                                  00003690
  803 FORMAT(/1X,22HU-VALUES FOR FUNCTION  I5,10H, STRATUM  I5)         00003700
  201 DO 20 K=1,NSAMP                                                   00003710
      SS = 0.                                                           00003720
      JC = 0                                                            00003730
      KY = KY + 1                                                       00003740
      DO 271 JD=1,NVD                                                   00003750
      NVC = HIRO(NVA,NV-JC)                                             00003760
      IF (IND2A .GT. 1) GO TO 2710                                      00003770
      READ(IND) W                                                       00003780
      GO TO 2711                                                        00003790
 2710 DO 2712 J=1,NVC                                                   00003800
      JC = JC + 1                                                       00003810
 2712 W(J) = W(KY,JC)                                                   00003820
 2711 IF (NVD .EQ. 1) GO TO 273                                         00003830
      READ(INY) YY                                                      00003840
      READ(IND) ID                                                      00003850
  273 DO 27 J=1,NVC                                                     00003860
      IF (ID(J) .EQ. 0) GO TO 27                                        00003870
      SS = SS + YY(J)*W(J)                                              00003880
   27 CONTINUE                                                          00003890
  271 CONTINUE                                                          00003900
      IF (JVAR .EQ. 0) GO TO 2713                                       00003910
      SS = SS/FRAC                                                      00003920
 2713 IF (NVD .EQ. 1) GO TO 274                                         00003930
      REWIND INY                                                        00003940
      REWIND IND                                                        00003950
  274 IF (JU .EQ. 0 .AND. KVAR .EQ. 0) GO TO 182                        00003960
      ICT = ICT + 1                                                     00003970
      U(ICT)=SS                                                         00003980
      IF (ICT .LT. KU .AND. K .LT. NSAMP) GO TO 18                      00003990
      IF (JU .EQ. 0) GO TO 181                                          00004000
  182 WRITE(6,805) (U(JCT),JCT=1,ICT)                                   00004010
  805 FORMAT(5(1X,E14.6))                                               00004020
```

```
101  IC1 = 0                                                          00004080
     IF (INNU .EQ. 0) GO TO 18                                        00004090
     WRITE(INU) U                                                     00004100
  18 IF (KVAR .NE. 0) GO TO 20                                        00004110
102  S = S + 55                                                       00004120
     IF (K .EQ. 1) GO TO 20                                           00004130
     S = K                                                            00004140
     Q = Q + (S-S-55)**2/(S*(S-1.))                                   00004150
  20 CONTINUE                                                         00004160
     IF (KVAR .NE. 0) GO TO 19                                        00004170
     BS = NSAMP                                                       00004180
     Q = (1. - FRAC)*BS*Q/(BS - 1.)                                   00004190
     GO TO 1                                                          00004200
  19 IF (KNU .EQ. 0) GO TO 191                                        00004210
     REWIND INU                                                       00004220
191  CALL VARNCE(U,KU,I,NSAMP,FRAC,Q,INU)                            00004230
     IF (KNU .EQ. 0) GO TO 1                                          00004240
     REWIND INU                                                       00004250
   1 A = A + Q                                                        00004260
     VAR = A                                                          00004270
     SAR = SQRT(A)                                                    00004280
     RVFR=VFR                                                         00004290
     CAR=SAR/PASS(VFR)                                                00004300
     IF (IDZA .GT. 1) GO TO 1020                                      00004310
     REWIND INU                                                       00004320
     IF (KS .EQ. 0) GO TO 1020                                        00004330
     REWIND INS                                                       00004340
1020 IF (JFVT .EQ. 2 .OR. KN .EQ. 0) GO TO 102                        00004350
     REWIND INF                                                       00004360
 102 WRITE(6,81) L,RVFR,VAR,SAR,CAR                                   00004370
  81 FORMAT(///1X,9HFUNCTION  I5,4(1X,E15.8))                         00004380
     RETURN                                                           00004390
     END                                                              00004400
     SUBROUTINE SWITCH(KA,KB)                                         00004410
     JA = KA                                                          00004420
     JB = KB                                                          00004430
     KA = JB                                                          00004440
     KB = JA                                                          00004450
     REWIND KA                                                        00004460
     REWIND KB                                                        00004470
   1 RETURN                                                           00004480
     END                                                              00004490
```